A fully automated tortuosity quantification system with application to corneal nerve fibres in confocal microscopy images

Annunziata, Roberto; Kheirkhah, Ahmad; Aggarwal, Shruti; Hamrah, Pedram; Trucco, Emanuele

Published in:
Medical Image Analysis

DOI:
10.1016/j.media.2016.04.006

Publication date:
2016

A Fully Automated Tortuosity Quantification System with Application to Corneal Nerve Fibres in Confocal Microscopy Images

Roberto Annunziata, Ahmad Kheirkhah, Shruti Aggarwal, Pedram Hamrah, Emanuele Trucco

PII: S1361-8415(16)30024-X
DOI: 10.1016/j.media.2016.04.006
Reference: MEDIMA 1102

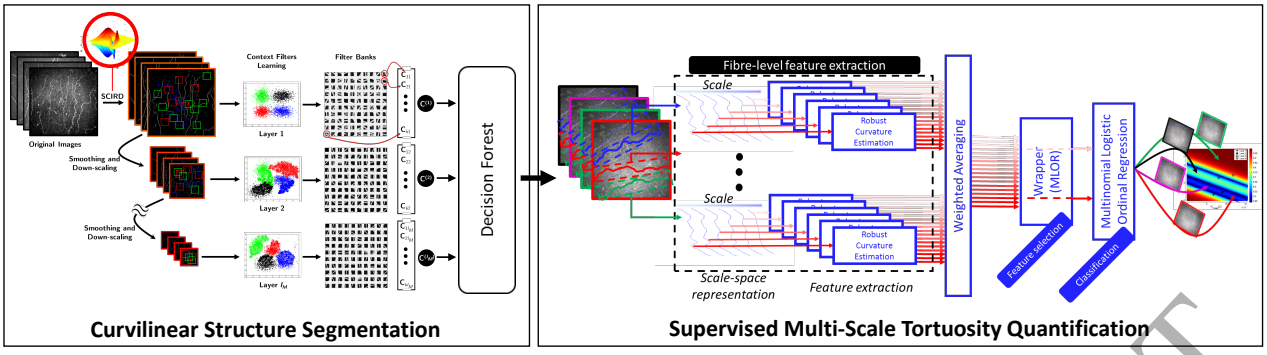
To appear in: *Medical Image Analysis*

Received date: 19 November 2015
Revised date: 20 April 2016
Accepted date: 20 April 2016

Please cite this article as: Roberto Annunziata, Ahmad Kheirkhah, Shruti Aggarwal, Pedram Hamrah, Emanuele Trucco, A Fully Automated Tortuosity Quantification System with Application to Corneal Nerve Fibres in Confocal Microscopy Images, *Medical Image Analysis* (2016), doi: 10.1016/j.media.2016.04.006

Highlights

- Data: 100 normal and pathological corneal images covering 4 tortuosity levels.
- Segmentation: combining a ridge detector with learned multi-range context filters.
- Multi-range context filters improve single-range context filters.
- Supervised, multi-scale tortuosity from highly discriminative measures.
- Rotation-invariant, spline-based fitting and analytic curvature estimation.



A Fully Automated Tortuosity Quantification System with Application to Corneal Nerve Fibres in Confocal Microscopy Images

Roberto Annunziata[a,*], Ahmad Kheirkhah[c], Shruti Aggarwal[c], Pedram Hamrah[b,c], Emanuele Trucco[a]

[a] *Computer Vision and Image Processing group, VAMPIRE project, School of Science and Engineering (Computing), University of Dundee, Dundee, UK*
[b] *Boston Image Reading Center and Cornea Service, New England Eye Center, Tufts Medical Center, Boston, USA*
[c] *Ocular Surface Imaging Center, Cornea Service, Massachusetts Eye & Ear Infirmary, Department of Ophthalmology, Harvard Medical School, Boston, USA*

Abstract

Recent clinical research has highlighted important links between a number of diseases and the tortuosity of curvilinear anatomical structures like corneal nerve fibres, suggesting that tortuosity changes might detect early stages of specific conditions. Currently, clinical studies are mainly based on subjective, visual assessment, with limited repeatability and inter-observer agreement. To address these problems, we propose a fully automated framework for image-level tortuosity estimation, consisting of a hybrid segmentation method and a highly adaptable, *definition-free* tortuosity estimation algorithm. The former combines an *appearance* model, based on a *Scale and Curvature-Invariant Ridge Detector* (SCIRD), with a *context* model, including multi-range learned context filters. The latter is based on a novel tortuosity estimation paradigm in which discriminative, *multi-scale* features can be automatically learned for specific anatomical objects and diseases. Experimental results on 140 *in vivo* confocal microscopy images of corneal nerve fibres from healthy and unhealthy subjects demonstrate the excellent performance of our method compared to state-of-the-art approaches and ground truth annotations from 3 expert observers.

Keywords: tortuosity, cornea, multiscale, segmentation, curvature, automated

1. Introduction

The tortuosity of curvilinear structures in the human body has received much attention since the seminal papers by Edington (1901) and Cairney (1924).

Numerous studies have reported correlations between several pathologies and the tortuosity of a wide range of anatomical structures such as retinal vessels (Heneghan et al., 2002; Cheung et al., 2011; Sasongko et al., 2011, 2015; Longmuir et al., 2010; Muraoka et al., 2014; Maude et al., 2014), intracerebral vessels (Bullitt et al., 2005) and conjunctival blood vessels Owen et al. (2008), but also the coronary (Eleid et al., 2014), iliac (Coulston et al., 2014), carotid (Bogunović et al., 2012) and aortic (Franken et al., 2015) arteries, the optic nerve (Ji et al., 2013) and corneal nerve fibres (Kallinikos et al., 2004; Hamrah et al., 2010; Kurbanyan et al., 2012; Hamrah et al., 2013; Edwards et al., 2014).

The pathologies investigated affect a large portion of the population worldwide and include diabetes, diabetic retinopathy and diabetic neuropathy (Sasongko et al., 2011, 2015; Edwards et al., 2014), retinopathy of prematurity (Heneghan et al., 2002; Wilson et al., 2008), malignant gliomas (Bullitt et al., 2004), facioscapulohumeral muscular dystrophy (Longmuir et al., 2010), spontaneous coronary artery dissection (Eleid et al., 2014), central retinal vein occlusion (Muraoka et al., 2014), children with neurofibromatosis type 1 (Ji et al., 2013). In particular, tortuosity has been investigated in corneal diseases such as unilateral herpes zoster (Hamrah et al., 2013), herpes simplex keratitis (Hamrah et al., 2010), acute acanthamoeba and fungal keratitis (Kurbanyan et al., 2012) and diabetic neuropathy (Kallinikos et al., 2004; Edwards et al., 2014).

In several studies (e.g., Hamrah et al., 2013; Eleid et al., 2014; Muraoka et al., 2014; Lagali et al., 2015), tortuosity was assessed by experienced specialists typically grading structures or whole images on a 3-5 level scale or as normal/abnormal based on qualitative, albeit structured, protocols (Oliveira-Soto and Efron, 2001). Regardless of the specific anatomical object of interest, such assessment is subjective and can lead to substantial inter-observer variability and possibly non-negligible intra-observer variability, thus limiting the sensitivity of the assessment scheme. Moreover, requiring direct inspection by specialists limits the amount of images that can be analysed and make large screening programs unfeasible or at least very expensive.

Several definitions of tortuosity have been proposed to try and quantify tortuosity automatically (Hart et al., 1999; Dougherty and Varro, 2000; Bullitt et al., 2003; Kallinikos et al., 2004; Wilson et al., 2008; Grisan et al., 2008; Trucco et al., 2010), but no single definition is widely accepted, possibly because tortuosity has different characteristics for different anatomical structures. We argue therefore that a highly adaptable tortuosity estimation algorithm, learning key features for specific image types and structures (*definition-free*) would be a promising and effective new research target.

However, efficiency would still be limited by manual seg-

*Corresponding author.
Email address: r.annunziata@dundee.ac.uk (Roberto Annunziata)

mentation, especially with today's image resolutions, e.g. $3,500 \times 2,300$ pixels for standard fundus camera images (Budai et al., 2009). Fast and accurate curvilinear structure segmentation is therefore needed, but different characteristics of tortuous curvilinear structures across image modalities makes segmentation challenging. In fact, tortuousity violates one of the basic assumptions of most tubular structure detectors, namely *locally straight tubular shape* (Frangi et al., 1998; Soares et al., 2006; Law and Chung, 2008; Hannink et al., 2014). Further issues include the presence of non-target structures (clutter), low resolution, noise and non-uniform illumination.

Fully automated tortuosity estimation frameworks have been proposed for retinal blood vessels, brain vasculature and corneal nerve fibres (Heneghan et al., 2002; Joshi et al., 2010; Scarpa et al., 2011; Koprowski et al., 2012). Typically, inaccuracies in the segmentation are the main source of inaccurate tortuosity estimates (e.g., Scarpa et al., 2011). Moreover, the aforementioned methods are based on mathematical definitions of tortuosity providing *fixed* models of a subjective perception. These measures combine features such as the number of inflection points along the structure, and the length to chord ratio. We argue that such hand-crafted tortuosity definitions limit the accuracy of an automated framework of general value.

1.1. Contributions

We propose a novel, fully automated framework for tortuosity estimation and demonstrate its performance with corneal nerve fibres in *in vivo* confocal microscopy (IVCM) images. Our system relies on a fast, robust and accurate segmentation algorithm and on a novel formalisation of tortuosity estimation. Validation is carried out both in terms of overall agreement with three experienced specialists, and of performance of each system module compared to several state-of-the-art methods. In addition, we quantify the accuracy in terms of tortuosity estimation when replacing manual segmentation (normally taken as ground truth for segmentation) with the proposed automated segmentation.

Our novel segmentation algorithm is specifically designed for tortuous and fragmented structures. It combines the benefits of an *appearance* model based on the *Scale and Curvature Invariant Ridge Detector* (SCIRD) and a *context* model, a multi-scale hierarchy based on learned context filters.

Finally, we propose a novel, adaptable framework for tortuosity estimation by learning flexibly important features for specific image modalities and structures. This is accomplished by tortuosity feature selection (FS) and supervised classification.

1.2. Related Work

1.2.1. Curvilinear structure segmentation

Many solutions have been proposed for curvilinear structure segmentation. Challenges include low signal-to-noise ratio at small scales, confounding non-target structures, non-uniform illumination and complex configurations (see Lesage et al. (2009) for an extensive review).

The extraction of adequate features is a fundamental aspect and relies mostly on the following three approaches.

(1) Hand-Crafted Features (HCFs). These methods are based on hand-designed filters modelling local geometrical properties of ideal tubular structures. Eigenvalue decomposition of the Hessian matrix was employed by Frangi et al. (1998); Santamaría-Pang et al. (2007); Martinez-Perez et al. (2007); Annunziata et al. (2015a); maximum projections over each scale were used to make the approach scale invariant. These projections were then used to build the well-known tubularity measure called *vesselness* by Frangi et al. (1998). However, performance tend to degrade at crossings or bifurcations since this approach only looks for elongated structures. To overcome this issue, Hannink et al. (2014) proposed to segment crossings/bifurcations with multiscale invertible orientation scores and apply vesselness filters to maps of the latter. Optimally Oriented Flux (OOF) was recently proposed by Law and Chung (2008) to improve detection of adjacent structures with vesselness measures. OOF is based on the computation of an optimal projection direction minimizing the inward oriented flux at the boundary of localized circles (2-D) or spheres (3-D) of different radii. This projected flux can be regarded as a likelihood that a pixel is part of a tubular structure. Tubularity measures can be obtained by combining the eigenvalues of the OOF Hessian matrix. Other successful HCFs rely on Gabor wavelets (aka Morlet wavelets) as proposed, for instance, by Soares et al. (2006) who exploited their directional selectiveness to detect oriented structures and achieve fine tuning to specific frequencies. HCFs have also been used by Honnorat et al. (2011) to compute a local tubularity measure feeding a graphical model. While HCFs are typically fast, their assumptions might be violated in real images by highly fragmented and tortuous structures, which limits detection performance and consequently the accuracy of tortuosity estimation.

(2) Fully Learned Architectures (FLA). These are designed to overcome HCFs modelling issues by learning object representations directly from training data, with excellent performance reported on several tasks (Farabet et al., 2013; Krizhevsky et al., 2012). FLA learn automatically the specific characteristics of the training set, such as inter-pixel dependencies. However, training complex learning architectures requires high-performance hardware and/or optimised implementations, and more importantly large datasets to avoid overfitting, which are not always available for medical images. For this reason, Becker et al. (2013) have recently proposed a less complex solution employing gradient boosting to learn convolutional filters and boosting weights simultaneously, applied with success to retinal blood vessels and neurites. Sironi et al. (2014) have used the responses of convolutional filters learned by sparse coding (henceforth, SC) as input features to multiple regressors trained to predict the distance from the centreline.

(3) Hybrid Methods (HM). This solution combines HCFs with filters learned by FLAs, exploiting the efficiency of fast HCFs while limiting the amount of learned filters. The first HM applied to tubular structures was proposed by Rigamonti and Lepetit (2012) and combines OOF with learned *appearance* filters, i.e. learned on the original training images. It employs convolutional SC to learn 9 appearance filters. Quantitative results show a clear improvement over methods based only on

HCFs, achieving the same level of performance obtained with a a filter bank of 121 filters learned via SC, at a limited computational cost. However, applying learned filters at the same *layer* where HCFs are then applied (i.e. original image patches) may result in redundant learned filters, potentially damaging the discrimination power of the feature set.

After feature extraction, or in combination with feature learning in the case of fully learned architecture, a supervised classifier is typically employed to find optimal decision boundaries in the feature space as, for instance, in Soares et al. (2006); Santamaría-Pang et al. (2007); Rigamonti and Lepetit (2012); Becker et al. (2013); Schneider et al. (2015).

1.2.2. Tortuosity estimation

Automated tortuosity quantification is particularly challenging due to the variability of anatomical objects and to the varying importance of each tortuosity feature (e.g., curvature, number of inflection points) for different structures of interest (e.g., retinal vessels, cononary artery) within specific clinical contexts (e.g., diabetes, malignant gliomas, facioscapulohumeral muscular dystrophy). Several definitions have been proposed with the aim of making tortuosity estimation objective (see Lisowska et al. (2014) for a comparative study on retinal blood vessels).

The distance measure (DM), or arch to chord length, has been widely used (e.g., Heneghan et al., 2002), mainly because of its simplicity. However, the DM tends to assign the same value to potentially very different shapes as it ignores inflection points. For this reason, Bullitt et al. (2003) proposed to multiply the DM by the number of inflection points along the vessels, and Grisan et al. (2008) proposed to sum the DM computed for each segment between inflection points and then multiply by the number of inflection points.

A different class of measures are integral functions (weighted sums) of curvature estimates along centrelines. Hart et al. (1999) compared seven such measures using two test problems, classifying blood vessel segments, and the whole vessel network as tortuous and non-tortuous. The sum of the squared curvatures along vessel centrelines was found to associate best with expert annotations for their data set.

Comparatively little work exists on the tortuosity of corneal fibres. Kallinikos et al. (2004) were the first to propose an objective, semi-automated method for quantifying sub-basal nerves tortuosity. Scarpa et al. (2011) adapted the algorithm by Grisan et al. (2008) for corneal nerve images, reporting good results with 30 images divided into 3 classes by a single ophthalmologist.

The measures above are based on a fixed, hand-crafted combination of measures, limiting adaptability to other anatomical objects and, arguably, to different pathologies showing different tortuosity characteristics.

1.3. Paper overview

The remainder of the paper is organised as follows. Section 2 describes the data set and annotations used for validation. The proposed system for automated tortuosity estimation is discussed in Section 3 and includes our solution to the curvilinear structure segmentation task (Section 3.1), the post-processing used to optimise segmented structures for tortuosity estimation (Section 3.2) and our approach to tortuosity estimation (Section 3.3). Parameter setting, guidelines and performance measures are reported in Section 4 for segmentation (Section 4.1) and tortuosity estimation (Section 4.2). Section 4.3 discusses how each module of the system was trained and tested. In Section 5 results are presented and discussed. Finally, we draw conclusions and suggest future research directions in Section 6.

This paper presents and extends our previous work (MICCAI workshop on ophthalmic medical image analysis 2014 (Annunziata et al., 2014), and MICCAI 2015 (Annunziata et al., 2015b,c)). Specifically, we employ SCIRD (Annunziata et al., 2015c) to model the *appearance* of tortuous and fragmented structures. Herein, we provide further details (e.g., the explicit function used for curved-support gradient estimation) and an improved mathematical formulation. Moreover, we propose to combine SCIRD with a modified, multi-range version of the learned *context* filters introduced in Annunziata et al. (2015b) and show improved performance. Tortuosity estimation follows the paradigm introduced by Annunziata et al. (2014), but a novel algorithm for accurate curvature estimation is introduced and shown to reduce processing time by several orders of magnitude, with the same or even better performance. Fully automated tortuosity estimation, our main goal, was not discussed in any of our previous papers.

2. Material

140 2-D images with 384 × 384 pixels were selected by the clinical authors from an existing database of images of the sub-basal nerve plexus in the central cornea, acquired with laser scanning *in vivo* confocal microscopy (IVCM) (Heidelberg Retina Tomograph 3 with the Rostock Cornea Module, Heidelberg Engineering GmbH, Heidelberg, Germany). The diode laser source of this microscope has a 670 *nm* red wavelength and the microscope is equipped with a 63× objective lens with a numerical aperture of 0.9 (Olympus, Tokyo, Japan). The images obtained by this confocal microscope represent a coronal section of the cornea of $400 \times 400\ \mu m^2$ which can be of any corneal layer.

For this general, albeit pilot-level study, images were selected from healthy and unhealthy subjects (49 women and 27 men, 54.9 ± 19.3 years old) with pathologies showing significant changes in terms of tortuosity (Figure 1). Specifically, 20 images were selected from healthy subjects, 20 images from patients with Herpes Simplex Virus (HSV) keratitis and 100 images from subjects with Dry Eye Disease (DED).

For ground truth, corneal nerve fibre centrelines were manually traced by a clinical author using NeuronJ[1], an add-on plug-in for the ImageJ software[2] (Cruzat et al., 2011). Tortuosity was determined using a clinically accepted grading scale (grades 1 through 4) (Oliveira-Soto and Efron, 2001) by three experienced observers (clinical authors) independently.

[1] `http://www.imagescience.org/meijering/software/neuronj`.
[2] `http://imagej.nih.gov/ij`.

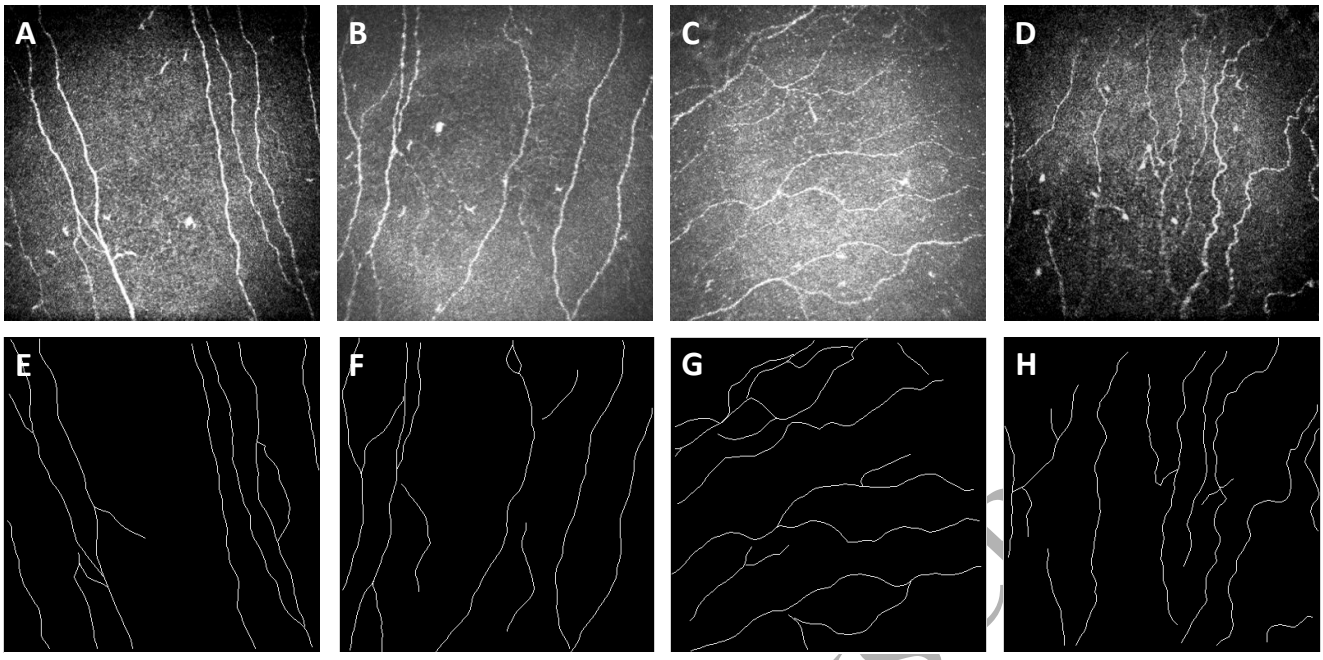


Figure 1: Four examples of corneal nerve fibre images captured through *in vivo* confocal microscopy. A,B,C,D: original images with increasing tortuosity level (A = 1, D = 4). E, F, G, H: manually traced fibre centrelines for image A, B, C, D, respectively.

Segmenting corneal nerve fibres in this data set is particularly challenging due to the presence of poorly contrasted, fragmented and highly tortuous fibres. Moreover, various images contain confounding, non-target structures such as dendritic cells easily mistaken for fibres given similar appearance. Finally, the clinical ordering in 4 tortuosity levels calls for higher discrimination capability compared to the previous fully automated solutions which were tested on 2 or 3 levels (e.g., Hart et al., 1999; Scarpa et al., 2011; Trucco et al., 2010).

3. Methods

3.1. Curvilinear structure segmentation

Appearance features model object-specific morphometric properties of curvilinear structures and can be hand-crafted (HCFs), or learned directly from original images. Such appearance features are used to enhance tube-like shapes (e.g., the longitudinal and cross-sectional intensity profiles). To leverage the information captured by the adopted appearance features and model inter-object relationships efficiently, we propose a multi-scale architecture based on learned *context* filters.

3.1.1. Appearance model: Scale and Curvature Invariant Ridge Detector

Our appearance model is designed specifically to detect tortuous and fragmented structures. To this end, we adopt our recently reported hand-crafted ridge detector, SCIRD, achieving curvature invariance by relaxing the standard assumption of locally straight tubular shape (Frangi et al., 1998; Soares et al., 2006; Law and Chung, 2008). To enforce connectivity, we augment the SCIRD filter bank with directionally elongated filters, similar to a Gabor filter bank Soares et al. (2006). First, we describe the adopted tortuous curvilinear structure model. Then, we derive the proposed ridge detector.

Curvilinear structure model. Our HCF is inspired by the curved-Gaussian model introduced by Lin and Dayan (1999) to analyse the correlation between currencies in time (exchange rate curvature analysis). Let $G(\boldsymbol{\varphi};\boldsymbol{\sigma})$ be a multivariate, n-D Gaussian function with diagonal covariance matrix, centred at the origin of the coordinate system,

$$G(\boldsymbol{\varphi};\boldsymbol{\sigma}) = \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^{n} \sigma_i^2}} \exp\left(-\sum_{i=1}^{n} \frac{\varphi_i^2}{2\sigma_i^2}\right) \tag{1}$$

where $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_n)$ represents a point in the $\{\varphi\}$ coordinate system, and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ describes the standard deviation in each direction. We bend the support of this Gaussian with a volume-preserving non-linear transformation $\mathcal{T} : \mathbb{R}^n \mapsto \mathbb{R}^n$ with $\mathcal{T}(\boldsymbol{x}) = \boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_n)$ of the form

$$\varphi_n = x_n + \sum_{i=0}^{n-1} k_{ni} m_{ni}(x_1, x_2, \ldots, x_{n-1}) \tag{2}$$

where k_{ni} are weights (see below) and the non-linear functions $m_{ni}(x_1, x_2, \ldots, x_{n-1})$ have continuous partial derivatives (Lin and Dayan, 1999). The change of variables theorem assures us that the non-linear transformation of the normalised $G(\boldsymbol{\varphi};\boldsymbol{\sigma})$ in the $\{\varphi\}$ coordinate system is already normalised in the new $\{x\}$ coordinate system since the transformation considered is volume-preserving.

In the 2-D case ($n = 2$), plugging Equation (2) in Equa-

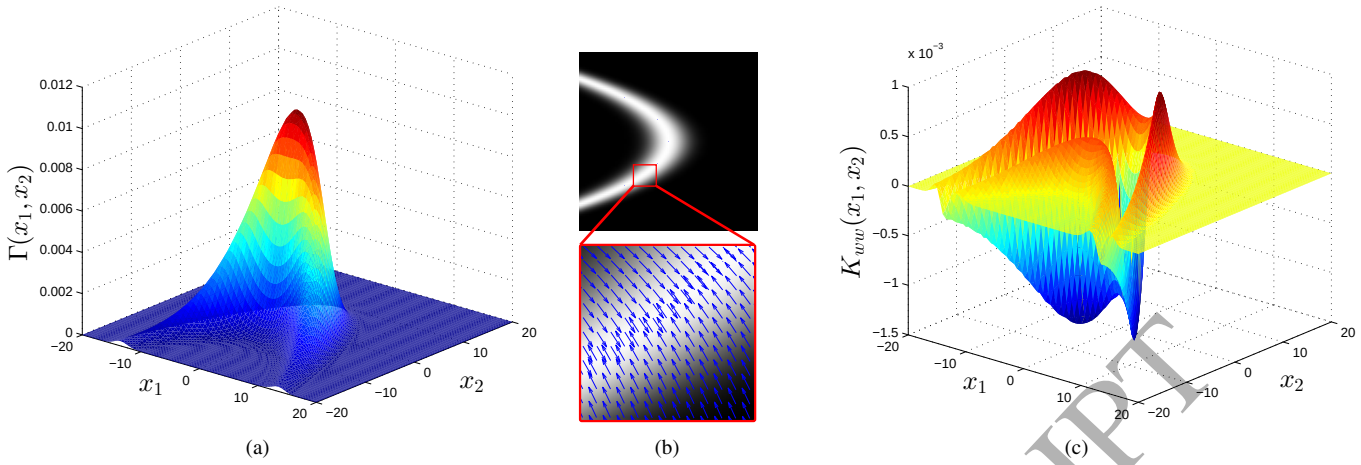


Figure 2: *Appearance* model used to capture specific characteristics of tortuous and fragmented structures. (a) An example of our curved-support shape model; (b) Gradient map used to compute the *tubularity probing filter* in the direction orthogonal to the model ideal centreline; (c) Derived convolutional filter estimating the local tubularity efficiently.

tion (1) leads to the curved-support bivariate Gaussian function:

$$G(x_1, x_2; \boldsymbol{\sigma}, \boldsymbol{k}) = \frac{1}{2\pi\sigma_1\sigma_2} \times \exp\left(-\frac{(x_1 + k_{10})^2}{2\sigma_1^2} - \frac{(x_2 + k_{20}m_{20} + k_{21}m_{21})^2}{2\sigma_2^2}\right), \quad (3)$$

where m_{20} and m_{21} depend on x_1.

If we now consider *quadratic* non-linear functions $m_{ni} = x_i^2$ for $0 < i < n$ and $m_{n0} = 1$, some of the parameters k_{ni} can be regarded intuitively as curvatures (Lin and Dayan, 1999). Specifically, we observe that k_{10} controls the elongation *asymmetry* of the shape (i.e. $k_{10} \neq 0$ makes one tail longer than the other), k_{21} its curvature and k_{20} is simply a translation parameter. However, although towards the end-point of a curvilinear structure the local shape would appear asymmetric longitudinally, we set $k_{10} = 0$ to keep the model simple and reduce the amount of free parameters to tune. This seems reasonable since tortuosity estimation, our reference application, would not benefit significantly from a particularly accurate segmentation of end-point regions. Finally, assuming that our model is centred on the specific curvilinear structure of interest, we set $k_{20} = 0$.

The model we adopt to detect tortuous and fragmented curvilinear structures is therefore:

$$\Gamma(x_1, x_2; \boldsymbol{\sigma}, k) = \frac{1}{2\pi\sigma_1\sigma_2} \underbrace{\exp\left(-\frac{x_1^2}{2\sigma_1^2}\right)}_{\text{longitudinal}} \underbrace{\exp\left(-\frac{\left(x_2 + kx_1^2\right)^2}{2\sigma_2^2}\right)}_{\text{orthogonal}}, \quad (4)$$

where (x_1, x_2) is a point in the principal component coordinate system of the target structure, (σ_1, σ_2) control the elongation of the shape ("memory") and its width, respectively; in fact, the first exponential of $\Gamma(x_1, x_2; \boldsymbol{\sigma}, k)$ controls the longitudinal Gaussian profile of the model, and the second the cross-sectional one. Unlike previous hand-designed models (e.g., Frangi et al., 1998; Soares et al., 2006; Law and Chung, 2008), we add a new parameter, k, to control the curvature of the Gaussian support (Figure 2(a)).

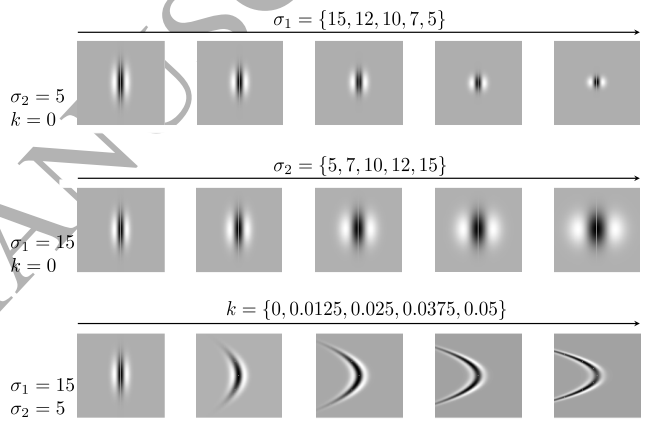


Figure 3: The effect of geometric parameters on the shape of SCIRD filters.

SCIRD. In order to estimate local tubularity, we adopt a shape-aware measure of contrast between the region inside and outside the curved ridge. Importantly, both the cross-sectional and the longitudinal Gaussian weighting are taken into account, allowing an accurate contrast estimation. In short, measuring the second directional derivatives along the cross-sectional profile of a curved-support Gaussian (Figure 2(b)), we obtain *probe* kernels such as the one shown in Figure 2(c). The shape of these probe kernels can be controlled through σ_1, σ_2 and k as shown in Figure 3. For a detailed derivation of SCIRD please refer to Appendix A[3].

3.1.2. *Context model: multi-range context filters*

As discussed in Section 1.2.1, feature extraction based on hybrid methods is particularly appealing as exploiting the effi-

[3]SCIRD MATLAB code is available at `http://staff.computing.dundee.ac.uk/rannunziata/code.html`.

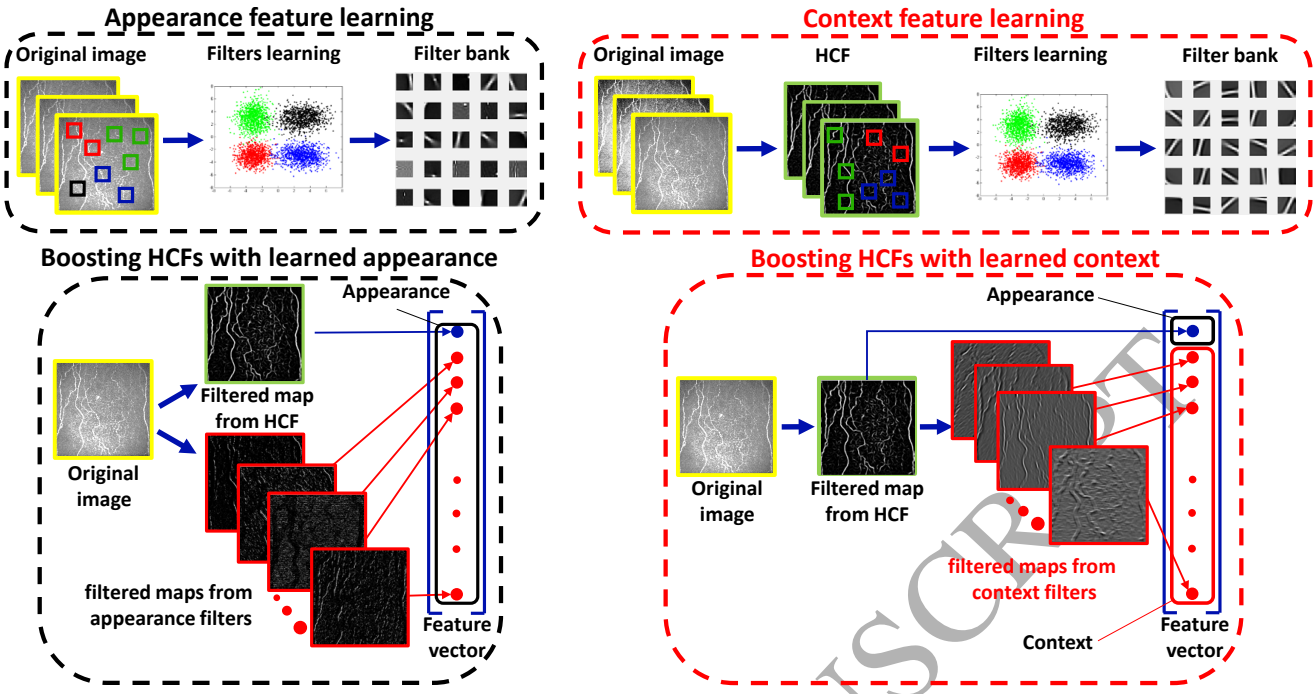


Figure 4: Difference in learning appearance and context filters. (Left) Appearance filters are learned using original image patches and applied on the same layer of the HCFs, thus leading to potential redundancy in the feature set. (Right) Context filters are learned using HCF maps and applied after the HCFs, thus eliminating redundancy. Notice, SCIRD is used as HCF here.

ciency of fast HCFs while limiting the amount of learned filters. Rigamonti and Lepetit (2012) proposed to combine learned *appearance* filters with HCFs *also* modelling *appearance* (i.e., OOF). However, learning inter-object relationships (i.e. *context*, as defined by Tu and Bai, 2010) and exploiting them in a segmentation framework has been recently proved to outperform significantly solutions based only on *appearance* features (e.g., Tu and Bai, 2010). We build on the same idea here.

Unsupervised filter learning. Although K-means is not designed to learn sparse representations, such as SC or independent component analysis, experimental results (Coates and Ng, 2012) suggest that it tends to learn sparse projections of the data under two assumptions: (1) sufficiently large number of training data, given the data dimensionality; (2) applying whitening to remove correlations between data components. Both are true in our case (Section 4.1), and we employ K-means clustering to learn context filters instead of more expensive algorithms such as SC (Rigamonti and Lepetit, 2012). We adopt the algorithm by Coates and Ng (2012) which we summarise concisely.

Our goal is to learn a dictionary $D \in \mathbb{R}^{q \times K}$ of K vectors so that a data vector $\mathbf{x}^{(i)} \in \mathbb{R}^q, i = 1, \ldots, m_D$ can be mapped to the code vector that minimizes the reconstruction error. Before running the learning algorithm we normalize the brightness and contrast of each input data point $\mathbf{x}^{(i)}$, in our case $p \times p$ patches. Then, we apply patch-level whitening through the ZCA transform so that $\mathbf{x}^{(i)}_{ZCA} = V(\Sigma + \epsilon_{ZCA}\mathbb{I})^{-1/2}V^\top \mathbf{x}^{(i)}$, where V and Σ are computed from the eigenvalue decomposition of the data points covariance $V\Sigma V^\top = \mathrm{cov}(X)$, and ϵ_{ZCA} is a small constant (Section 4.1) controlling the trade-off between whitening and noise amplification. After pre-processing the patches, we solve the optimization problem

$$\underset{D,\mathbf{c}}{\operatorname{argmin}} \sum_i \left\| D\mathbf{c}^{(i)} - \mathbf{x}^{(i)}_{ZCA} \right\|_2^2 \tag{5}$$

subject to $\|\mathbf{c}^{(i)}\|_0 \leq 1, \forall i = 1, \ldots, m_D$ and $\|\mathbf{d}^{(j)}\|_2 = 1, \forall j = 1, \ldots, K$, where $\mathbf{c}^{(i)}$ is the code vector related to input $\mathbf{x}^{(i)}_{ZCA}$, and $\mathbf{d}^{(j)}$ is the j-th column of the dictionary D.

Combining appearance filters learned through K-means (or SC) with HCFs leads to the combination approach proposed in Rigamonti and Lepetit (2012) based on appearance-only features (see Figure 4 left).

Unsupervised context filters learning. Hand-designing context filters to capture *inter*-object relationships is particularly challenging as the configurations to be considered, especially in the medical domain, can vary significantly. To overcome this issue, Tu and Bai (2010) proposed *auto-context*, a framework that learns multiple discriminative models sequentially. Auto-context achieves excellent performance on several object segmentation tasks (Tu and Bai, 2010), but with an extra computational cost compared to methods based on a single discriminative model. This may have little impact on training/testing time for applications involving small data sets or small images, but it can become impractical otherwise, for instance, with large screening programs like diabetic retinopathy (Ikram et al., 2013).

We therefore aim to include context information without increasing computational cost with respect to the hybrid solution proposed by Rigamonti and Lepetit (2012) based on a single discriminative appearance model. We achieve this by learning a *single* classifier (e.g., Decision Forest) which takes as input

both *appearance* (estimated by SCIRD) and *context* information.

Learning context filters has two advantages over appearance filters: 1) including inter-object relationships in a hybrid framework; 2) eliminating, or reducing significantly, the redundancy in the combination of HCFs and learned filters. In fact, learned appearance filters may be reconstructed through a combination (linear or non-linear) of the HCFs already used to model appearance, thus reducing the discrimination power of the feature set. Figure 4 shows the difference between the proposed approach (right) and the combination method proposed by Rigamonti and Lepetit (2012) (left). Notice, while appearance filters are learned on the original image (where HCF are applied), context filters are learned *after*.

Unsupervised multi-range context filters learning. The approach above does capture some level of context, but it has limitations. First, the *range* of spatial context captured in each direction (p) is dictated by the patch size ($p \times p$). This parameter can be set by cross-validation, but its maximum value p_M is limited by the amount of available training data (see assumption (1) above on applying K-means). Notice that the data points dimensionality $q \sim p^2$, thus forcing p_M to be relatively small (e.g., less than 30 pixels) if training data is not abundant (e.g., less than 500,000 patches), a typical case in medical applications. So, this approach fails to model *long-range* context information. Second, it is a *single-range* context model, i.e. it captures inter-object relationships characterising a specific neighbourhood only. We address these issues by introducing a multi-scale architecture of learned context filters modelling *multi*-range spatial context.

Let us denote with $\{I_t^{(S_1)}\}$ the set of tubularity maps obtained after SCIRD is applied to the training images $\{I_T\}$, where $T = 1, \ldots, T_M$. The proposed multi-range context filters architecture is obtained by learning filters on $\{I_T^{(S_1)}\}$ and their smoothed, downscaled versions $\{I_T^{(S_\mathcal{L})}\}$, where $\mathcal{L} = 1, \ldots, l_M$ indicates a certain level of this architecture. As Figure 5 shows, at level $\mathcal{L} = 1$ patches randomly sampled from $\{I_T^{(S_1)}\}$ are used to learn the first set of K context filters $C^{(1)} = [C_{11}|\ldots|C_{K1}]^\top$, where C_{ij} is $p \times p$ pixels. This captures context information in the range of p pixels in each direction. Then, Gaussian smoothing with σ_{MRC} and downscaling with factor 2 (i.e., image width and height are halved) using bi-cubic interpolation are applied to the tubularity maps $\{I^{(S_1)}\}$ to obtain $\{I^{(S_2)}\}$ and access the level $\mathcal{L} = 2$. Context filter learning is applied to these downscaled images to learn a second set of K filters $C^{(2)} = [C_{12}|\ldots|C_{K2}]^\top$, C_{ij} is again $p \times p$ pixels. This process is applied repeatedly until level $\mathcal{L} = l_M$ is reached and the respective set of context filters $C^{(l_M)} = [C_{1l_M}|\ldots|C_{Kl_M}]^\top$ is learned.

Notice, context filters size is the same (i.e., $p \times p$) at each level $\mathcal{L}$ but images width and height are halved from $\mathcal{L} = i$ to $\mathcal{L} = i + 1$, doubling the context range in each direction. This makes our context learning solution *multi*-range. Moreover, although the number of available training patches reduce as images are downscaled, K-means can be still employed with good clustering performance, making the proposed solution fast and efficient.

3.1.3. Training and Testing

The first step is to apply SCIRD to each original image $I \in \{I_T\}$ to obtain the tubularity map $I_T^{(S_1)}$. The value of the tubularity map at a certain pixel, $SCIRD(x, y)$, corresponds to the first component of the feature vector $\mathbf{f}(x, y)$ representing this pixel. Then, the correlation between each of the K context filters learned at the level $\mathcal{L} = 1$ and SCIRD response is computed at each pixel (x, y) to obtain K *context* feature maps $\left[I_{11}^{(C)}|\ldots|I_{K1}^{(C)}\right]^\top$, where

$$I_{n1}^{(C)}(x, y) = \sum_{j=-\frac{p}{2}}^{\frac{p}{2}} \sum_{i=-\frac{p}{2}}^{\frac{p}{2}} SCIRD(x + i, y + j) C_{n1}(i, j), \qquad (6)$$

$\forall n = 1, 2, \ldots, K$ and C_{n1} is the n-th context filter arranged as $p \times p$ matrix.

At level $\mathcal{L} = 2$, each full size SCIRD response $I^{(S_1)} \in \{I_T^{(S_1)}\}$ is smoothed and downscaled. Then, second-level learned context filters $C^{(2)} = [C_{12}|\ldots|C_{K2}]^\top$ are applied to build the second set of K context feature maps $\left[I_{12}^{(C)}|\ldots|I_{K2}^{(C)}\right]^\top$. Upscaling is then applied to recover full image size. This process is repeated until level $\mathcal{L} = l_M$ is reached and the l_M-th set of K context feature maps is computed. In the end, each pixel is represented with both *appearance* and *multi*-range *context* features:

$$\mathbf{f} = \Big[\underbrace{SCIRD}_{appearance}, \underbrace{I_{12}^{(C)}, \ldots, I_{K2}^{(C)}}_{context,\ \mathcal{L}=1}, \ldots, \underbrace{I_{1l_M}^{(C)}, \ldots, I_{Kl_M}^{(C)}}_{context,\ \mathcal{L}=l_M} \Big]^\top. \qquad (7)$$

A random subset of positive and negative instances represented by the feature vectors $\{\mathbf{f}\}$ is sampled by each training image and used to train a random decision forest (henceforth, RF) with N_T trees and maximum number of samples in each leaf N_l (see Section 4 for a discussion on the values of these parameters and their choice).

During testing, the probability $P(\ell = 1 \mid \mathbf{f})\ \forall (x, y) \in \mathcal{D}(I)$ with label $\ell \in \{0, 1\}$ is estimated by feeding the trained RF classifier with the computed feature vector $\mathbf{f}$.

Centrelines are the key information to assess tortuosity (e.g., Scarpa et al., 2011; Sasongko et al., 2011). The role of caliber is not yet fully understood in this context (Trucco et al., 2010). In our approach (and for all the baselines) we apply fast Canny-like non-maxima suppression (henceforth, NMS) to the tubularity maps obtained.

3.2. Post-processing for tortuosity estimation

After NMS, we obtain a centreline map. To get the final segmentation for a test image, we apply the global threshold T_{pm} maximising the F-measure on the training partition of a leave-one-out strategy (see examples in Figure 6). Finally, all the isolated segments whose length is below a certain threshold T_{pp} are removed. The threshold is estimated using cross-validation.

Ophthalmologists typically focus on specific structures (e.g., major retinal vessels at a specific distance from the optic disc, Sasongko et al. (2015)). Small branches are generally not taken into account when tortuosity is assessed at image level. To obtain isolated, long structures suitable for automated tortuosity estimation we employ a simple, yet fast pruning algorithm.

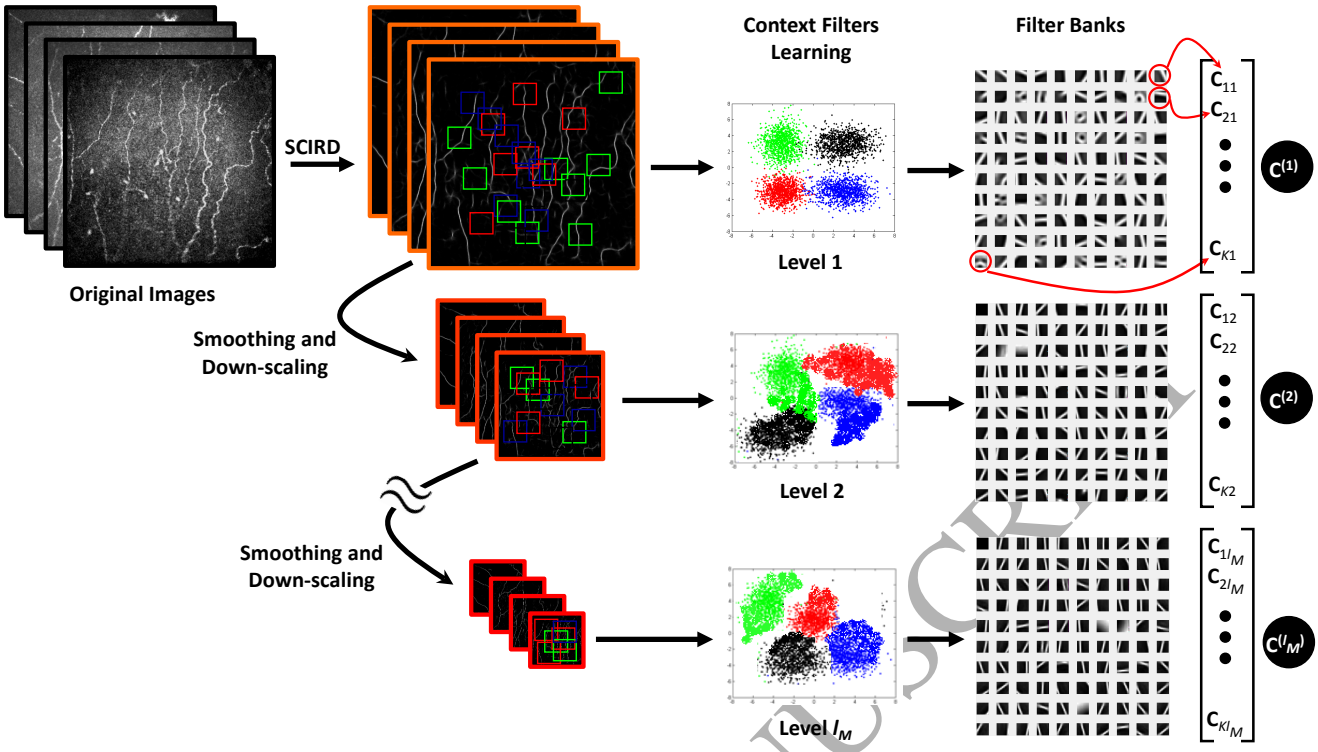


Figure 5: Block diagram of the proposed unsupervised multi-range context filters learning. Notice, we keep the patch size constant over the levels to capture larger and larger spatial context (light orange, orange and red indicate short-, medium- and long-range context, respectively).

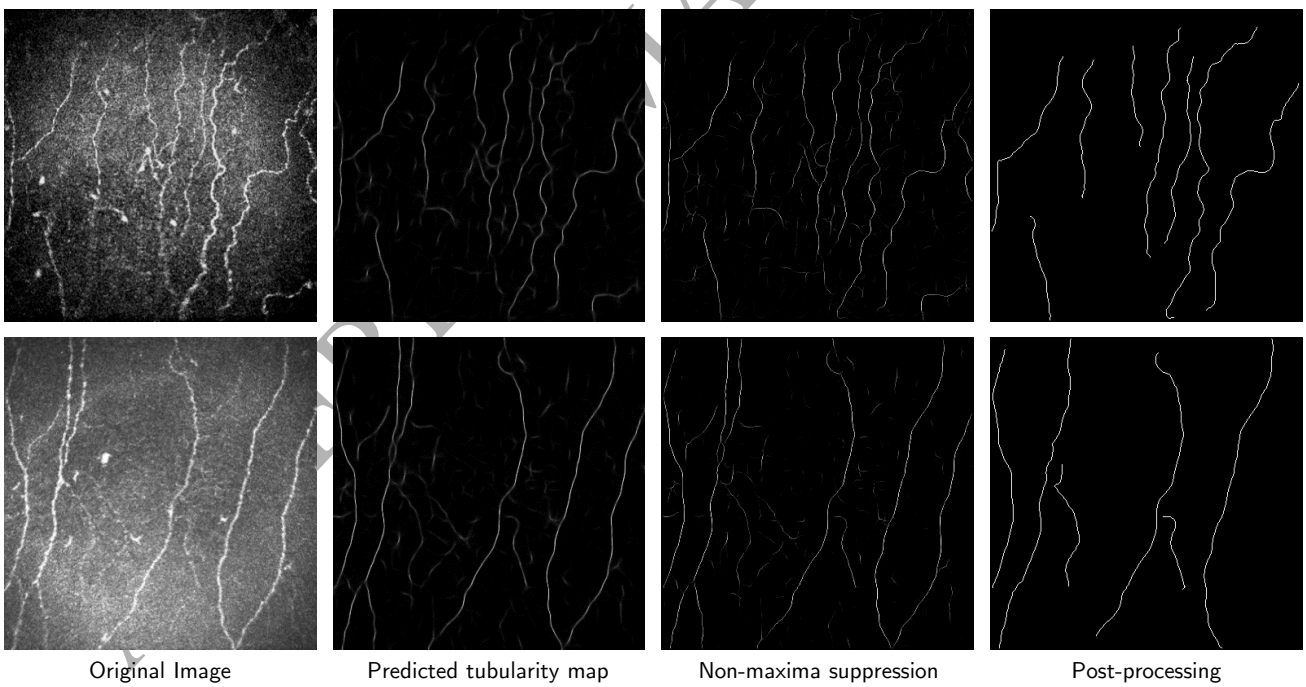


Figure 6: Visual examples of our segmentation pipeline for two very diverse cases in our data set.

This selects the longest path connecting two endpoints, for each tree-like connected component within the image. First, the Euclidean distance is measured among all the endpoints of a connected component, and the farthest points are selected. Second, the path minimising the geodesic distance is considered the main structure of interest for that specific connected component (see examples in Figure 6).

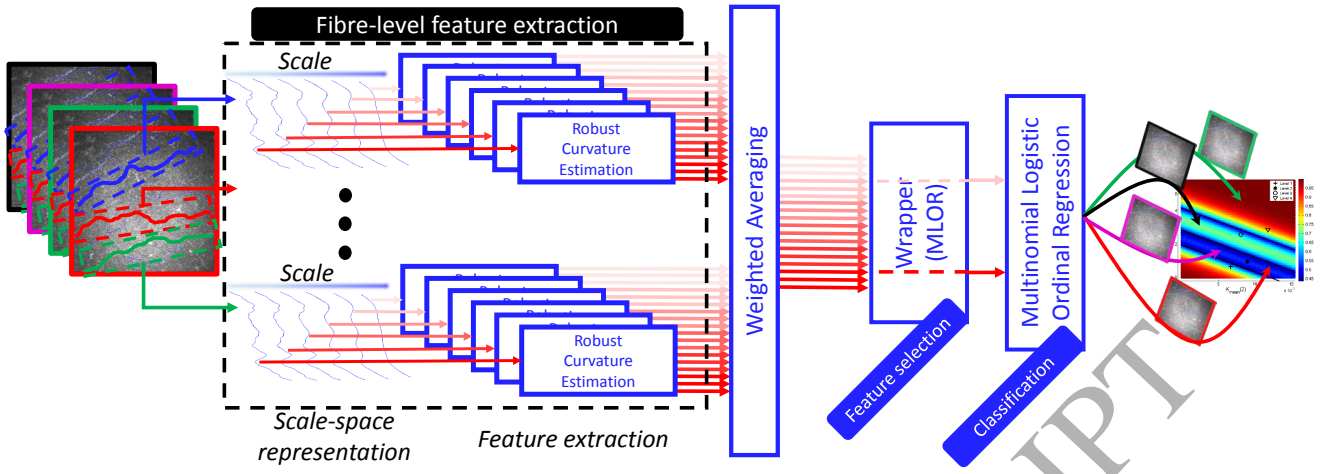


Figure 7: Block diagram of our tortuosity estimation framework: wrapper-based feature selection followed by multinomial logistic ordinal regression. First, for every corneal nerve fibre in a image, robust spline fitting is applied for fast and accurate curvature estimation at each level of the fibre's scale-space representation. Mean curvature $k_{mean}(t)$, twistedness $d_{ip}(t)$, and maximum curvature $k_M(t)$ are computed for each scale $t \in \{1, \ldots, t_M\}$ so that a feature vector $v_f = \{k_{mean}(1), \ldots, k_{mean}(t_M), d_{ip}(1), \ldots, d_{ip}(t_M), k_M(1), \ldots, k_M(t_M)\}$ is obtained. Second, weighted (fibre length is used as weight) averaging v_f across all the fibres creates the pool of image-level features $v_{img} = \{K_{mean}(1), \ldots, K_{mean}(t_M), D_{ip}(1), \ldots, D_{ip}(t_M), K_M(1), \ldots, K_M(t_M)\}$. Third, a wrapper-based FS technique is employed for identifying the most discriminative combination of features and their scale. Finally, a multinomial logistic ordinal regressor is used to assign image-level tortuosity.

3.3. Tortuosity estimation

Tortuosity estimation is a particularly difficult task as different anatomical structures in the context of different pathologies can show different tortuosity characteristics. Most of the previously reported methods lack adaptability as they are based on fixed, *postulated* combinations of tortuosity features. Moreover, different frequencies of direction changes are ignored, while an in-depth visual investigation of corneal images, for instance, suggests that a multiscale approach for tortuosity features extraction is indeed needed. Finally, curvature estimation algorithms are typically based on finite differences leading to noisy estimates, especially when segmentations are obtained automatically.

Our method addresses the issues above. It can be summarised in four steps (Figure 7):

1. compute multi-scale shape features for each curvilinear structure to obtain its tortuosity representation;
2. combine shape features computed from each curvilinear structure to obtain image-level features;
3. use FS to identify the most discriminative set of image-level features, which become the tortuosity representation *for that specific application*;
4. use the set of features identified to train a regressor assigning new images to one of a tortuosity grade on a fixed scale.

3.3.1. Multi-scale tortuosity representation of a single curvilinear structure

Curvature-based measures tend to outperform the combination of DM (i.e., maximum deviation from chords) and number of inflection points, when high sampling rates are used to represent vessel centrelines (Lisowska et al. (2014); see also Section 1.2.2).

We showed previously (Annunziata et al., 2014) that curvature-based measures, if estimated accurately, can lead to accuracy comparable or even higher than that of two experienced observers, when the third is taken as reference. Our algorithm for curvature estimation was based on a multi-window ellipse/line fitting approach which can be time consuming, although sub-sampling could achieve a reasonable trade-off between performance and time.

Here, we use spline fitting to replace our multi-window ellipse/line fitting for fast and accurate curvature estimation. The structures of interest are normally rotated first or represented in parametric coordinates. We choose the latter to avoid additional computational burden and potential inaccuracies introduced by rotations. To counteract the noise introduced by the automated centreline segmentation, we adopt a robust cubic spline fitting solution[4]. This yields accurate curvature estimates, supporting reliable estimation of inflection points.

Derivatives are computed locally and *analytically* from the fitted splines. Changes in curvature sign along the curves are the estimated inflection points (Figure 8).

In summary, the following shape features are extracted from each fibre (Annunziata et al., 2014):

- the average curvature along the fibre, k_{mean};
- the "twistedness", or density of inflection points, d_{ip};
- the maximum curvature along the fibre, k_M.

We propose to represent curvilinear structures for tortuosity estimation using a scale-space representation, taking into account the different frequency of direction changes. The multi-scale

[4]MATLAB code available at `http://www.mathworks.com/matlabcentral/fileexchange/13812-splinefit`

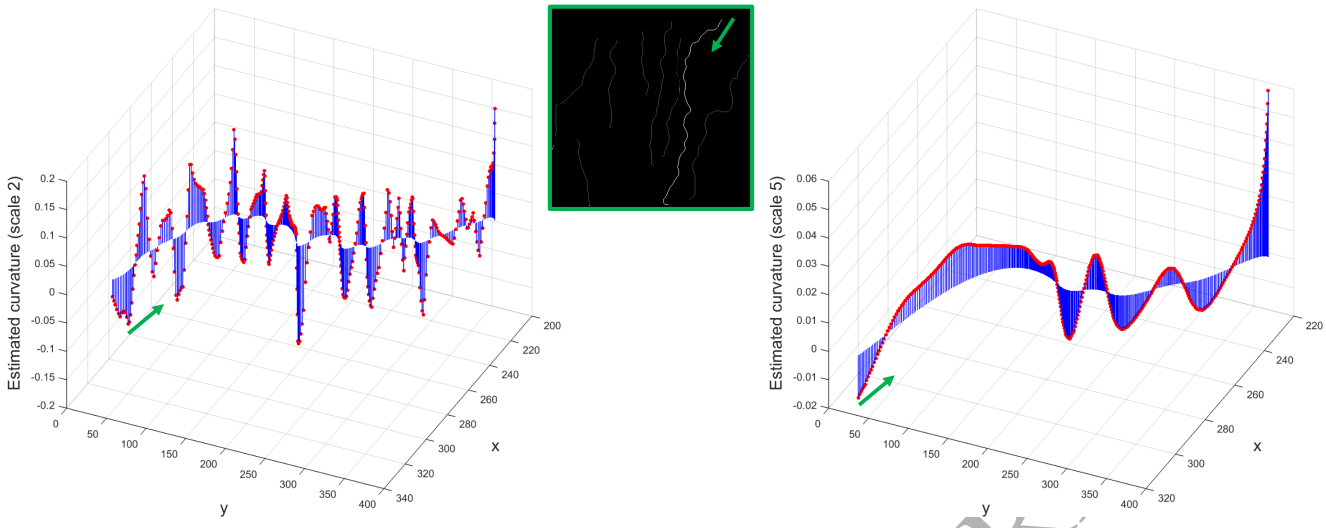


Figure 8: Accurate multi-scale curvature estimation via cubic spline fitting. Stem plots illustrate the estimated local curvature for the automatically segmented corneal nerve fibre shown in the top image. Left, right: estimated curvature at scale 2 and 5, respectively. The green arrows indicate starting point ($l = 0$) and direction ($l > 0$).

version of the features listed above becomes: $\{k_{mean}(t)\}$, $\{d_{ip}(t)\}$ and $\{k_M(t)\}$, for $t \in \{1, \ldots, t_M\}$, where t_M is the coarsest resolution used.

3.3.2. *Image-level tortuosity features for image classification*

Our next task is *to assign a tortuosity grade to a whole image*. This is nontrivial as images contain variable numbers of corneal nerve fibres of varying lengths, which could show considerably different tortuosity characteristics. We turn the features computed for individual fibres into *image-level features* by computing, at each scale t and for each feature, a weighted average over all fibers in the image, where the weights are the fibre lengths. The weights model the observation that longer fibres are more informative, overall, than shorter ones.

Denoting the length of the $i-th$ fibre at scale t as $l_i(t)$, and the total number of fibres within an image as N_f, our image-level features are defined as follows:

- $K_{mean}(t) = \frac{\sum_{i=1}^{N_f} l_i(t) k_{mean}(t)}{\sum_{i=1}^{N_f} l_i(t)}$;
- $D_{ip}(t) = \frac{\sum_{i=1}^{N_f} l_i(t) d_{ip}(t)}{\sum_{i=1}^{N_f} l_i(t)}$;
- $K_M(t) = \frac{\sum_{i=1}^{N_f} l_i(t) k_M(t)}{\sum_{i=1}^{N_f} l_i(t)}$.

3.3.3. *Feature selection and tortuosity prediction*

FS identifies the most discriminative features and the most important spatial scales for the tortuosity assessment. This choice makes our tortuosity estimation framework highly adaptable.

Of the several FS strategies proposed (Peng et al., 2005; Lazar et al., 2012; Guyon and Elisseeff, 2003), we adopt a wrapper-based strategy in which a learning algorithm is used to score repeatedly subsets of features according to their predictive power. Exhaustive search can be employed as our feature pool is relatively small (i.e. less than 20, including different spatial scales), guaranteeing that the search returns the true absolute extremum. We use a multinomial logistic ordinal regressor (henceforth, MLOR), a multi-class (low-complexity, hence fast) classifier taking into account the order of tortuosity grades (1 to 4, i.e., ranking), modelled by a *logit* link function. Specifically, MLOR models the *log cumulative odds* (Bishop et al., 2006), the logarithm of the ratio of the probability of a class preceding or equal to class j in the class ranked list, $P(y_c \leq c_j)$ and the probability of the same class following class j, $P(y_c > c_j)$. To limit model complexity and make FS more efficient, we assume that the effects of features f_i are the same for all classes on the logarithmic scale. This means that the model has different intercepts (γ), but common slopes (β) among classes (proportional odds assumption). Thus, the MLOR model is

$$\ln\left(\frac{P(y_c \leq c_j)}{P(y_c > c_j)}\right) = \gamma_j + \beta_1 f_1 + \beta_2 f_2 + \cdots + \beta_s f_s \qquad (8)$$

for $j \in \{1, 2, 3\}$, where s is the feature space cardinality.

Once the most discriminative combination of features is selected by exhaustive search, these are extracted from training images and a new MLOR model is learned and used to predict unseen images.

4. Experimental Setup and Performance Evaluation

Table 1 shows the complete list of parameters we adopted or learned to train and test the proposed tortuosity estimation system. We provide explanations and guidelines to facilitate adaptation to other data sets, potentially including different curvilinear structures.

4.1. *Curvilinear structure segmentation*

We manually tuned SCIRD parameters σ_1, σ_2, k controlling the filter elongation, width and curvature, respectively. The

Table 1: Parameters of the proposed automated tortuosity estimation system. Refer to Section 4 for guidelines related to each parameter setting.

System parametrisation

Segmentation

Parameter	Adopted value	Description
σ_1	{2,3,4}	SCIRD elongation.
σ_2	{2,3}	SCIRD width.
k	{0,0.025,..., 0.1}	SCIRD curvature.
θ	{15, 30, ..., 180}	SCIRD rotation angles.
α	1	SCIRD contrast enhancement.
$p \times p$	15×15 pixels	Context filters size.
ϵ_{ZCA}	0.001	Patch-level whitening in Spherical K-means (Coates and Ng, 2012).
K	100	Number of context filters learned for each pyramid level.
$\mathcal{L}$	3	Number of pyramid levels set manually in order to incorporate multi-range context.
σ_{MRC}	1	Sigma for Gaussian smoothing when learning multi-range context filters.
N_T	100	Number of trees in the decision forest.
N_l	Learned	Maximum number of samples in each leaf of the decision forest.
T_{pm}	Learned	Threshold on the probability map obtained as output of the decision forest.
T_{pp}	Learned	Threshold on the fibre's length used in the post-processing.

Tortuosity Estimation

Parameter	Adopted value	Description and guidelines
t_M	6	Number of scales used to build the scale-space representation of each fibre.
γ_i, β_i	Learned	Coefficients of the MLOR model.

range of σ_1 was chosen considering the level of fragmentation in our data set (i.e. longer kernels were used to detect highly fragmented structures). To set σ_2 values, we took into account the maximum and minimum width of the target structures. Curvature values were set according to the variety of tortuosity levels shown by corneal nerve fibres in our data set. We set the rotation step of SCIRD filters to achieve a good compromise between running time and segmentation accuracy. The contrast normalisation parameter α was set as discussed in Section 3.1.1. Notice that the number of SCIRD free parameters and the experience required to set them is in line with other HCFs (e.g., Gabor, Frangi, OOF). However, parameters for SCIRD and baseline HCFs were tuned separately to achieve their best performance and provide a fair comparison.

We learned 100 15×15 filters using at least $100,000$ patches for each context level. Increasing the number of context filters is expected to improve segmentation performance, at the price of a slower tortuosity estimation. Our experiments suggest that 100 filters per level represent a good compromise between training/testing speed and tortuosity estimation performance. Moreover, we tested the sensitivity of the segmentation performance to the filter size and found that performance are practically unchanged when $p = \{11, 15, 21\}$ pixels. When setting ϵ_{ZCA} we considered the compromise between maximising filter sharpness and limiting noise amplification, as suggested by Coates and Ng (2012). We set σ_{MRC} and $\mathcal{L}$ in our multi-range context architecture to capture short-, medium- and long-range context. In particular, σ_{MRC} controls the level of detail preserved in each context level, while $\mathcal{L}$ controls how long is the range of context to be captured (i.e. the number of context levels). According to Criminisi et al. (2012) increasing N_T should lead to better classification performance, at the cost of slower predictions. We found experimentally that $N_T = 100$ is enough for our application. The depth of the forest is automatically learned using the out-of-bag error on the training set.

We compare our segmentation methods with the recent hybrid solution proposed by Rigamonti and Lepetit (2012) as well as with other baselines, including widely used and state-of-the-art ridge detectors: Frangi (Frangi et al., 1998), based on Hessian matrix eigenvalue analysis, Gabor filtes (Soares et al., 2006) and the recent Optimally Oriented Flux (OOF) (Law and Chung, 2008).

Following the recent literature, including Rigamonti and Lepetit (2012), we adopt precision-recall curves to assess centreline detection performance, and use use a tolerance factor ρ (Sironi et al., 2014): a predicted centreline point is considered a true positive if at most ρ pixels away from the closest ground truth centreline point. In our experiments we set $\rho = 2$ pixels. Following the established benchmarking protocol (Rigamonti and Lepetit, 2012; Sironi et al., 2014), we average performance measures over 5 random sub-sampling cross-validation runs, using 70 images for training and the rest for testing in each run.

Thresholds on the probability maps obtained as output of the decision forest (i.e. T_{pm}) and on the fibre's length used in the post-processing (i.e. T_{pp}) are learned as discussed in Section 4.3.

4.2. Tortuosity estimation

We compute image-level tortuosity measures (mean, max and density of inflection points) in a scale-space representation of 6 spatial scales (i.e. $t_M = 6$), therefore assessing the discriminative power of 18 features in total.

Following the standard performance assessment protocol for multi-class classification (Sokolova and Lapalme, 2009), we use *weighted* accuracy (Acc), sensitivity (Se) and specificity (Sp), positive predicted value (Ppv) and negative predictive value (Npv), defined below.

$$\mathrm{Acc} = \sum_{i=1}^{N_c} w_i \, \frac{\mathrm{TP}_i + \mathrm{TN}_i}{\mathrm{TP}_i + \mathrm{FN}_i + \mathrm{FP}_i + \mathrm{TN}_i} \tag{9}$$

$$\mathrm{Se} = \sum_{i=1}^{N_c} w_i \, \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{10}$$

$$\mathrm{Sp} = \sum_{i=1}^{N_c} w_i \, \frac{\mathrm{TN}_i}{\mathrm{TN}_i + \mathrm{FP}_i} \tag{11}$$

$$\mathrm{Ppv} = \sum_{i=1}^{N_c} w_i \, \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i} \tag{12}$$

$$\mathrm{Npv} = \sum_{i=1}^{N_c} w_i \, \frac{\mathrm{TN}_i}{\mathrm{TN}_i + \mathrm{FN}_i} \tag{13}$$

where TP_i indicates true positives, TN_i true negatives, FP_i false positives, FN_i false negatives for the $i-th$ tortuosity level; w_i represents the percentage of images whose GT level is i, according to observer taken as reference, as suggested by Sokolova and Lapalme (2009) ($N_c = 4$). The above performance measures do not take into account the distance from the actual value when considering errors. Therefore we also include the mean squared error (MSE) and mean absolute error (MAE) defined as

$$\mathrm{MSE} = \frac{1}{N_{im}} \sum_{i=1}^{N_{im}} (y_{c_i} - c_i)^2 \tag{14}$$

$$\mathrm{MAE} = \frac{1}{N_{im}} \sum_{i=1}^{N_{im}} \left| y_{c_i} - c_i \right| \tag{15}$$

where N_{im} is the number of test images, y_{c_i} is the predicted tortuosity level, and c_i is the true one for the $i-th$ image.

We compare the performance of our multi-scale tortuosity estimation approach with single-scale methods reported to perform very well in comparative tests: tortuosity density (TD Grisan et al., 2008), distance measure (DM Heneghan et al., 2002), slope chain coding (SCC Bribiesca, 2013), τ_5 (Hart et al., 1999). Moreover, we report our recent results obtained using a tortuosity estimation algorithm based on multi-window curvature estimation (Annunziata et al., 2014).

Since our data set was annotated by three experienced observers independently (clinical authors, reported as Obs_1, Obs_2, Obs_3), we take the classification by one observer in turn as ground truth, and compare the performance of ours and aforementioned methods with the other two observers. Therefore, MLOR coefficients are learned using a leave-one-out cross-validation for each observer taken as reference. Notice that 11 images were excluded (only for the tortuosity estimation performance evaluation, not for the segmentation) due to complete disagreement among the observers, confirming the difficulty of tortuosity modelling.

We investigate the impact of each building block in the proposed framework on the tortuosity estimation performance. Specifically, (1) we compare the performance of our tortuosity estimation algorithm with state-of-the-art methods, at a parity of segmentation; (2) we assess the performance when replacing a conventional segmentation method (i.e. based on the "locally-straight" assumption) with the proposed segmentation approach, at a parity of tortuosity estimation algorithm.

4.3. Evaluating each module individually

Segmentation. For each training partition (including 70 images) of each random subsampling run (5 in total) we repeat the following procedure. First, we learn the set of context filters by applying SCIRD and unsupervised context filter learning to the training partition, as described in Sections 3.1.1 and 3.1.2. Second, we apply SCIRD and the learned multi-range context filters to all the training images to build the feature vectors for each pixel, as described in Section 3.1.3. Third, we train a decision forest using a subset of the training feature vectors (i.e., we randomly sample 500 positive and 500 negative instances for each image). During this step, we learn the depth of the trees using out-of-bag error on the training set. Once the forest is trained, we apply SCIRD and the learned multi-range context filters to the test images of the current partition. For each testing partition we measure segmentation performance after non-maxima suppression, as described in Section 4.1.

Tortuosity Estimation. In order to obtain segmented centrelines suitable for tortuosity estimation on the whole data set, we split it randomly in two subsets. We apply the same strategy discussed above to train the segmentation module on each of the two training partitions and use it to obtain the segmentation (probability) maps on the testing partitions, in turn. We then apply non-maxima suppression to each image. To obtain the final segmentation suitable to the tortuosity analysis, the best thresholds on the probability map (i.e., T_{pm}) is learned using a cross-validation strategy. Specifically, T_{pm} is set as the threshold maximising the F-measure on each training set of this two-fold split and applied to the test set, in turn. At this point, tortuosity feature selection can be applied to identify the most discriminative set of features by using a wrapper strategy based on the MLOR model, as described in Annunziata et al. (2014). We adopt the majority voting ground truth for feature selection as we aim at discovering a "general" set of discriminative tortuosity features (i.e. valid among observers, not specific to each observer) using machine learning, an emerging need in the community (Lagali et al., 2015). Once the most discriminative set of tortuosity features is identified, we apply a leave-one-out strategy to learn a MLOR model on each training partition using the tortuosity GT levels provided by the observer taken as reference, and use this model to estimate the tortuosity level of the image held out. We iterate until the tortuosity of all the images has been quantified. Notice that the value of T_{pp} (i.e. the length on isolated connected components to discard), is learned on the training partition of the leave-one-out strategy used to learn the MLOR parameters and used for each image held out.

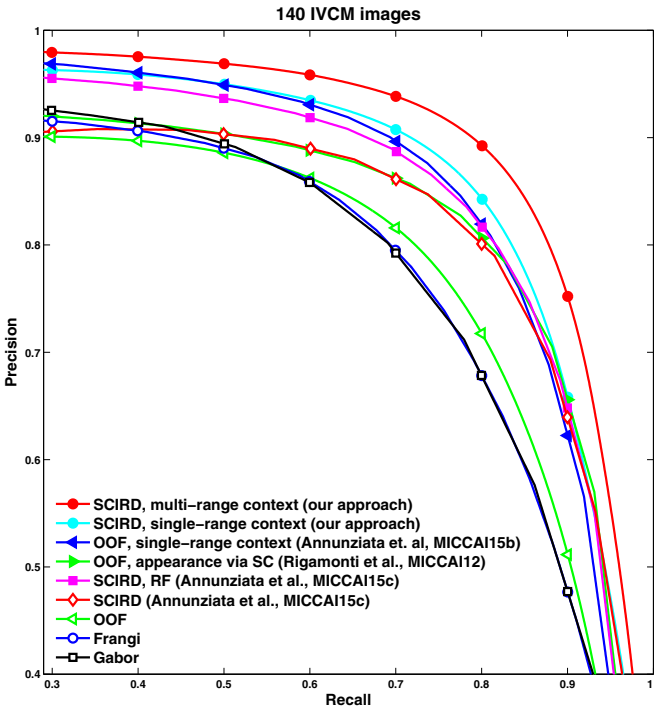


Figure 9: Precision-recall curves for the corneal nerve fibre centreline detection task (pixel-level classification). Each curve is obtained by averaging the results of 5 random cross-validation trials in which the whole data set, including 140 images, is randomly split in 2 equal partitions.

5. Results and Discussion

Curvilinear structure segmentation. Figure 9 shows the performance of the proposed segmentation method and the baselines, in terms of precision-recall curves. The comparison among HCFs shows that SCIRD performs best on our data set, therefore justifying our choice of adopting SCIRD as base HCF on top of which we learn multi-range context filters. Using SCIRD filter bank responses as input to a RF, i.e. "SCIRD, RF" (Annunziata et al., 2015c), leads to a significant improvement compared to plain SCIRD, especially at low-medium recall rates. Combining HCFs with learned (single-range) context filters leads to better segmentation results. However, using SCIRD instead OOF in the combination, yields better performance. A noticeable improvement is achieved by employing multi-range context filters, i.e. "SCIRD, multi-range context", showing the benefit of learning a discriminative model capturing context at different levels. Notice that the post-processing step described in Section 3.2 is not used to refine the segmentation results of any method reported in this figure, as it is only needed to adapt segmentation results to tortuosity quantification.

Qualitative comparisons with immediate competitors shown in Figure 10 suggest that improved performance is largely due to our multi-range context model which reduces significantly the amount of false positives far from target structures and improves the connectivity of corneal nerve fibres appearing fragmented. Nevertheless, our segmentation pipeline is fast as it exploits context information efficiently: since a single discriminative model (i.e., a single RF) is employed, a whole corneal nerve image is segmented in about 30 seconds, using unoptimized MATLAB code on a machine equipped with Intel i7-4770 CPU at 3.4 GHz. We carry out FS on our pool of multi-scale features using the majority voting ground truth for tortuosity estimation as reference[5]. We found through cross-validation that the best combination of image-level tortuosity features is $\{K_{mean}(2), K_{mean}(5)\}$, suggesting that a multi-scale approach is more suitable to tortuosity estimation than single-scale indices reported in literature. We adopted this combination of features in our experiments.

Comparison of tortuosity estimation algorithms at a parity of segmentation. Table 2 shows our experimental evaluation for the tortuosity estimation task when state-of-the-art tortuosity estimation approaches are used together with the best segmentation approach (i.e., "SCIRD, multi-range context", as shown in Figure 9). First, experimental results adopting each expert observer as reference in turn, suggest that the proposed tortuosity estimation method based on multi-scale cubic splines (MSSPLINE) outperforms the multi-scale-multi-window approach (MSMW) previously proposed (Annunziata et al., 2014). Moreover, MSSPLINE is much faster than MSMW as shown in Table 6, thus successfully addressing the main drawback of MSMW. Second, MSSPLINE performs significantly better than the baselines for all the observers used as reference, thus suggesting that a multi-dimensional tortuosity representation feeding a MLOR model is more suitable for this task. Third, tortuosity FS described in Section 3.3.3 yields an substantial extra performance when Obs_1 and Obs_3 are used as reference. This can be justified by the fact that some features may be highly correlated, noisy due to segmentation imperfections or not discriminative enough. Finally, the proposed fully automated tortuosity quantification system (i.e., "SCIRD, multi-scale context + MSSPLINE", reported in this table as MSSPLINE) matches or exceeds the tortuosity estimation performance of at least one of the other two expert observers, when compared against the one taken as reference.

Comparison of segmentation algorithms at a parity of tortuosity quantification approach. Table 3 shows the tortuosity estimation performance when replacing a conventional segmentation method (Rigamonti and Lepetit, 2012) with the proposed one. The tortuosity estimation algorithm used here is our MSSPLINE, shown to outperform other methods in Table 2. Experimental results suggest that modelling tortuous structures through SCIRD and using multi-range context filters leads to a significant improvement over conventional segmentation methods[6]. This improvement is particularly remarkable as, for the first time to our best knowledge, we show that it closes the gap between automated and manual segmentation for the tortuosity estimation task (see Table 2).

[5]The majority voting ground truth assumes that the true tortuosity level of each image is the one assigned by the majority of experienced observers, independently.

[6]Notice that we employ the post-processing described in Section 3.2 for both the automated segmentation methods (i.e. the one we propose and the one proposed by Rigamonti and Lepetit (2012)), for a fair comparison.

Table 2: **Tortuosity estimation: comparison of tortuosity estimation algorithms at a parity of segmentation approach.** For tortuosity ground truth, each experienced observer is taken as reference in turn (Obs_1, Obs_2 and Obs_3, clinical authors). The segmentation approach used for these experiments is our *"SCIRD, multi-range context"* shown to outperform others methods on our data set (see Figure 9). The first two columns in each table show the performance of the others observers against the one used as reference. The other columns in each table report several tortuosity estimation algorithms. Specifically, *MSMW* is the algorithm based on multi-scale-multi-window tortuosity (Annunziata et al., 2014), *MSSPLINE* is the proposed approach based on multi-scale rotation invariant spline fitting (results in boldface), *MSSPLINE (no FS)* is the latter approach without using tortuosity feature selection, TD is the tortuosity density index (Grisan et al., 2008), *DM* is the distance measure (Heneghan et al., 2002), SCC is the tortuosity estimation algorithm based on slope chain coding (Bribiesca, 2013) and τ_5 is the normalised integral of the squared curvature (Hart et al., 1999).

Ground Truth = Obs_1

Performance measure	Obs_2	Obs_3	MSMW	MSSPLINE	MSSPLINE (no FS)	TD	DM	SCC	τ_5
Acc	0.7678	0.7451	0.708	**0.7408**	0.7066	0.6247	0.6236	0.6194	0.6303
Se	0.5736	0.5194	0.4729	**0.5194**	0.4651	0.3333	0.3411	0.3333	0.2171
Sp	0.8457	0.8219	0.7825	**0.8113**	0.7904	0.6797	0.6732	0.6704	0.8083
PPv	0.5729	0.5547	0.4815	**0.5184**	0.463	0.1165	0.3842	0.1129	0.4864
Npv	0.8329	0.8223	0.7949	**0.8204**	0.7945	0.8379	0.841	0.8388	0.7346
MSE	0.4729	0.5736	0.6202	**0.5504**	0.6512	1.1705	1.0543	1.0853	2.6977
MAE	0.4419	0.5116	0.5581	**0.5039**	0.5736	0.8295	0.7907	0.8062	1.3333

Ground Truth = Obs_2

Performance measure	Obs_1	Obs_3	MSMW	MSSPLINE	MSSPLINE (no FS)	TD	DM	SCC	τ_5
Acc	0.7791	0.7513	0.7619	**0.7838**	0.7802	0.6304	0.6301	0.626	0.6661
Se	0.5736	0.5271	0.5426	**0.5814**	0.5736	0.3101	0.3101	0.3023	0.3101
Sp	0.8455	0.8254	0.8334	**0.8478**	0.8506	0.7055	0.7024	0.6997	0.8044
PPv	0.5977	0.558	0.5473	**0.5807**	0.5678	0.2278	0.3564	0.0928	0.5479
Npv	0.8507	0.8317	0.8386	**0.856**	0.8514	0.7864	0.837	0.835	0.7774
MSE	0.4729	0.5659	0.6202	**0.5349**	0.5426	1.2946	1.2946	1.3411	2.6124
MAE	0.4419	0.5039	0.5116	**0.4574**	0.4651	0.8915	0.8915	0.907	1.2326

Ground Truth = Obs_3

Performance measure	Obs_2	Obs_1	MSMW	MSSPLINE	MSSPLINE (no FS)	TD	DM	SCC	τ_5
Acc	0.7329	0.7381	0.6735	**0.7266**	0.6777	0.6366	0.6411	0.6388	0.6232
Se	0.5271	0.5194	0.4419	**0.5271**	0.4419	0.4109	0.4186	0.4109	0.2868
Sp	0.829	0.8275	0.6994	**0.7409**	0.7299	0.6008	0.5983	0.5969	0.7848
PPv	0.5495	0.5459	0.4223	**0.5057**	0.4293	0.2892	0.3187	0.1715	0.3269
Npv	0.793	0.8001	0.7711	**0.8286**	0.7617	0.7405	0.8778	0.8766	0.7306
MSE	0.5659	0.5736	0.7907	**0.6124**	0.7442	1.1008	1.1163	1.1473	2.2171
MAE	0.5039	0.5116	0.6357	**0.5194**	0.6202	0.7597	0.7597	0.7752	1.1473

Confusion matrices and Cohen's kappa. We further assess the performance of the proposed fully automated tortuosity quantification system by reporting confusion matrices and Cohen's kappa values when each of the three observers is used in turn as reference. As Table 4 shows, most of the images classified incorrectly are classified in the immediate following or preceding tortuosity level, a desirable property of our system, since tortuosity represents an ordinal variable. Results in terms of Cohen's kappa and modified kappa (i.e. $\mathbf{K}/\mathbf{K_M}$, where $\mathbf{K_M}$ is the maximum possible $\mathbf{K}$ given marginal frequencies (Dunn, 1989)) are reported in Table 5. Notice that given the different marginal frequencies for each tortuosity level (whose values change depending on the observer used as ground truth), the modified kappa represents a fairer index to assess agreement (Dunn, 1989). Nevertheless, we report the original kappa values for completeness. In both cases, the proposed system matches or exceeds the performance obtained by the other expert observers, when compared to the one taken as reference.

Running time. We report running time for the tortuosity estimation algorithm alone and used in combination with automated segmentation in Table 6: about 70 minutes are required to estimate image-level tortuosity for 140 images (fully automated). Thank to the adopted spline-based curvature estimation, the running time is reduced considerably. The current parameter setting (Section 4) seems to represent a good compromise between speed and tortuosity estimation accuracy.

Applying the system to other tortuosity quantification problems. The proposed system was designed and developed to work on tortuous structures in general, and could potentially be applied to structures other than corneal nerve fibres. Specifically, SCIRD and the single-range context model were found particularly suitable for other curvilinear structures as well, e.g. neurites acquired with different modalities, therefore presenting different characteristics in terms of contrast, resolution and context(Annunziata et al., 2015c). The multi-range context model introduced here improved performance significantly on corneal

nerve fibres; such improvements could be observed in similar scenarios, presenting non-target structures with same appearance characteristics or low signal-to-noise ratio. The approach for automated tortuosity quantification is highly versatile, as it is capable of identifying the most discriminative tortuosity features which may vary for different structures and, more importantly, for different pathologies.

We maintained the segmentation and tortuosity estimation modules separate to allow an easier and more effective interpretation of the results, compared to solutions based on direct tortuosity feature extraction and image classification. In our experience, the main source of errors in tortuosity estimation is the segmentation module, but training our hybrid segmentation solution requires a very limited amount of images and annotations, compared to fully-learned architectures. These choices should allow easier adaptation within the medical imaging domain, in which image data and annotations are not always abundant.

A limitation of the proposed framework is that SCIRD was derived and developed for 2-D images. An extension to 3-D is possible, but the number of filters could become prohibitively large. In this case, a faster solution could be obtained by using OOF or Frangi's method as base HCF.

6. Conclusions

Several studies have reported correlations between various diseases and the tortuosity of anatomical curvilinear structures. Such studies are often based on time-consuming, manual annotations and subjective, visual assessments, thus reducing repeatability and inter-observer agreement. We have addressed these problems by proposing a fully automated framework for image-level tortuosity estimation with application to corneal nerve fibres in IVCM.

Our fully automated tortuosity estimation system includes two modules: segmentation and tortuosity estimation. The former relies on a hybrid segmentation method combining an appearance model, based on a scale and curvature invariant ridge detector (SCIRD), with a context model, including multi-range learned context filters. The latter is based on a novel tortuosity estimation paradigm, capable of identifying the most discriminative tortuosity features and their combination, among the pool of multi-scale tortuosity measures (i.e. it does not pre-define which features determine tortuosity estimation).

Experiments carried out on 140 images from healthy and unhealthy subjects with different pathologies, show that our segmentation module for tortuous structures outperforms conventional methods based on the "locally-straight" assumption and learned appearance filters. The improvement in terms of segmentation is immediately transferred to tortuosity estimation performance. The tortuosity estimation module, based on multi-scale spline-based curvature estimation, performs significantly better than state-of-the-art *single*-index algorithms and addresses the main drawback of the previously proposed multi-scale-multi-window approach, i.e. the speed. In fact, the proposed solution is orders of magnitude faster than the previous one, and achieves better performance. This speed gain, combined with the hybrid segmentation solution, makes the system fast: 30s are required to analyse an IVCM image using a stan-

Table 3: **Tortuosity estimation: comparison of segmentation algorithms at a parity of tortuosity quantification approach.** For tortuosity ground truth, each experienced observer is taken as reference in turn (Obs_1, Obs_2 and Obs_3, clinical authors). The tortuosity quantification approach used for these experiments is our *MSSPLINE* shown to outperform other methods on our data set (see Table 2). The first two columns in each table show the performance of the others observers against the one used as reference. The other columns in each table report tortuosity quantification performance when the following approaches are used: manual segmentation (*Manual*); the proposed *"SCIRD, multi-range context"* segmentation algorithm (*Proposed*, in bold face), based on curved-support and multi-range context filters; the segmentation method proposed by Rigamonti and Lepetit (2012), based on "locally-straight" and appearance filters.

Ground Truth = Obs_1

Performance measure	Obs_2	Obs_3	Manual	Proposed	Rigamonti et al.
Acc	0.7678	0.7451	0.7448	**0.7408**	0.7182
Se	0.5736	0.5194	0.5426	**0.5194**	0.4806
Sp	0.8457	0.8219	0.8068	**0.8113**	0.7888
PPv	0.5729	0.5547	0.559	**0.5184**	0.4745
Npv	0.8329	0.8223	0.824	**0.8204**	0.8064
MSE	0.4729	0.5736	0.5736	**0.5504**	0.6589
MAE	0.4419	0.5116	0.4961	**0.5039**	0.5659

Ground Truth = Obs_2

Performance measure	Obs_1	Obs_3	Manual	Proposed	Rigamonti et al.
Acc	0.7791	0.7513	0.8039	**0.7838**	0.7327
Se	0.5736	0.5271	0.6202	**0.5814**	0.4806
Sp	0.8455	0.8254	0.8618	**0.8478**	0.8111
PPv	0.5977	0.558	0.6262	**0.5807**	0.4715
Npv	0.8507	0.8317	0.8696	**0.856**	0.8217
MSE	0.4729	0.5659	0.4729	**0.5349**	0.7829
MAE	0.4419	0.5039	0.4109	**0.4574**	0.5969

Ground Truth = Obs_3

Performance measure	Obs_2	Obs_1	Manual	Proposed	Rigamonti et al.
Acc	0.7329	0.7381	0.6828	**0.7266**	0.6589
Se	0.5271	0.5194	0.4496	**0.5271**	0.4031
Sp	0.829	0.8275	0.7055	**0.7409**	0.6811
PPv	0.5495	0.5459	0.4025	**0.5057**	0.2946
Npv	0.793	0.8001	0.7858	**0.8286**	0.7737
MSE	0.5659	0.5736	0.6899	**0.6124**	0.8217
MAE	0.5039	0.5116	0.5969	**0.5194**	0.6667

Table 4: Confusion matrices for the tortuosity estimation task obtained using the proposed method. For tortuosity ground truth, individual observers (clinical authors) are used in turn (Obs_1, Obs_2 and Obs_3).

		Estimated level			
		1	**2**	**3**	**4**
Obs_1 level	**1**	22	12	1	0
	2	12	26	4	1
	3	0	14	12	8
	4	0	1	9	7

		Estimated level			
		1	**2**	**3**	**4**
Obs_2 level	**1**	23	9	2	0
	2	10	25	3	1
	3	0	13	10	7
	4	0	2	7	17

		Estimated level			
		1	**2**	**3**	**4**
Obs_3 level	**1**	1	16	2	0
	2	1	45	5	2
	3	0	20	8	6
	4	0	2	7	14

Table 5: Cohen's kappa and modified kappa for the tortuosity estimation task. For ground truth, individual observers (i.e., clinical authors) are taken as reference in turn (Obs_1, Obs_2 and Obs_3). We compare the performance of the proposed method (i.e., *SCIRD, multi-range context + MSSPLINE*, indicated as *Ours* here) with the expert observers, in terms of Cohen's kappa (**K**, first row) and modified kappa ($\mathbf{K/K_M}$, second row).

	GT = Obs_1			**GT = Obs_2**			**GT = Obs_3**		
	Obs_2	**Obs_3**	**Ours**	**Obs_1**	**Obs_3**	**Ours**	**Obs_1**	**Obs_2**	**Ours**
K	0.42	0.34	**0.33**	0.42	0.36	**0.43**	0.34	0.36	**0.28**
$\mathbf{K/K_M}$	0.47	0.41	**0.37**	0.47	0.44	**0.48**	0.41	0.44	**0.43**

Table 6: Comparison in terms of running time for each tortuosity estimation algorithm alone and in combination with the automated segmentation algorithm proposed herein. The proposed spline-based curvature estimation, i.e. *MSSPLINE* (results in boldface), is significantly faster than our previous multi-window solution, i.e. *MSMW*. Experiments were carried out on Intel i7-4770 CPU @ 3.4 GHz, using MATLAB code. Each image was 384×384 pixels.

Execution Time (140 images)	**MSMW**	**MSSPLINE**	**TD**	**DM**	**SCC**	τ_5
Tortuosity (s)	7590	**23.1**	47.1	43	40.1	39
Segm. + Tortuosity (min)	196.5	**70.4**	70.8	70.7	70.7	70.7

dard computer with MATLAB. The comparison with three experienced observers who annotated the images independently shows, remarkably, that the proposed system matches or even exceeds their performance. This opens the possibility of analyzing the large volumes of images needed for screening programs very efficiently, subject of course to further validation with much larger data sets.

To allow an easy adaptation to tortuosity estimation problems involving curvilinear structures other than corneal nerve fibres, we have provided a detailed discussion about our system parametrisation and guidelines to set its parameters.

Next, we plan to deploy the proposed framework in large studies investigating corneal nerve fibre tortuosity and its relationship with specific pathologies.

Acknowledgment

R. Annunziata is supported by the EU Marie Curie Initial Training Network (ITN) "REtinal VAscular Modelling, Measurement And Diagnosis" (REVAMMAD), project number 316990. P. Hamrah is supported by the National Eye Institute Grant R01EY022695, Falk Medical Research Trust and Research to Prevent Career Development Award. We thank the VAMPIRE project members for useful discussions. We are also grateful to the anonymous reviewers for their valuable comments and suggestions, which contributed to improve this paper.

Appendix A. SCIRD derivation

Let $I(x, y)$ be the intensity of a monochromatic image at location (x, y) in image coordinates (the same derivation could be applied to each channel of a colour image). To make the contrast measure shape-aware locally, we introduce local first-order gauge coordinates $\{\mathbf{v}(x, y), \mathbf{w}(x, y)\}$, where $\mathbf{w}(x, y) = \frac{\nabla I(x,y)}{\|\nabla I(x,y)\|}$ and $\mathbf{v}(x, y) = \mathbf{w}_{\perp}(x, y)$.

A shape-aware contrast measure is then given by the second directional derivative in the direction orthogonal to the model centreline:

$$I_{\mathbf{ww}} = D_{\mathbf{w}}\left[D_{\mathbf{w}} I\right] = D_{\mathbf{w}}\left[\mathbf{w}^{\top}\nabla I\right] \triangleq \mathbf{w}^{\top} H_I \mathbf{w}, \tag{A.1}$$

where $D_{\mathbf{w}}$ is the directional derivative operator along $\mathbf{w}$.

$$H_I = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{bmatrix} \tag{A.2}$$

is the Hessian matrix, and $\nabla I = [I_x, I_y]^{\top}$ is the gradient of I in x-y coordinates. Substituting Equation (A.2) in Equation (A.1) we obtain:

$$I_{\mathbf{ww}} = \frac{(I_x I_{xx} + I_y I_{yx}) I_x + (I_x I_{xy} + I_y I_{yy}) I_y}{I_x^2 + I_y^2}. \tag{A.3}$$

where we have omitted arguments $(x, y; \boldsymbol{\sigma}, k)$ for compactness. We differentiate by convolving the image with derivatives of the curved-support Gaussian, which leads to an efficient tubularity estimation based on the convolution with a filter bank that *can be pre-computed off-line*:

$$I_{\mathbf{ww}}(x, y; \boldsymbol{\sigma}, k) = I(x, y) * K_{\mathbf{ww}}(x, y; \boldsymbol{\sigma}, k), \tag{A.4}$$

where $K_{\mathbf{ww}}$ represents our *tubularity probe kernel* (see example in Figure 2(c)):

$$K_{\mathbf{ww}} = \frac{(\tilde{\Gamma}_x \Gamma_{xx} + \tilde{\Gamma}_y \Gamma_{yx})\tilde{\Gamma}_x + (\tilde{\Gamma}_x \Gamma_{xy} + \tilde{\Gamma}_y \Gamma_{yy})\tilde{\Gamma}_y}{\tilde{\Gamma}_x^2 + \tilde{\Gamma}_y^2}. \tag{A.5}$$

Here, $\tilde{\Gamma}(x_1, x_2; \boldsymbol{\sigma}, k)$ is a curved-support Gaussian model with a constant (i.e. non-Gaussian) longitudinal profile, whose gradient direction is orthogonal to the centreline (Figure 2(b)):

$$\tilde{\Gamma}(x_1, x_2; \boldsymbol{\sigma}, k) \propto \exp\left(-\frac{\left(x_2 + k x_1^2\right)^2}{2\sigma_2^2}\right). \tag{A.6}$$

To achieve invariance in the discrete domain, we create a filter bank of $K_{\mathbf{ww}}(x, y; \sigma_1, \sigma_2, k, \theta)$ kernels generated (*off-line*) by making σ_2 (i.e., scale) and k (i.e., curvature) span the respective ranges, $\{\sigma_2^{(i)},\ i = 1, \ldots, N_{\sigma_2}\}$ and $\{k^{(i)},\ i = 1, \ldots, N_k\}$, in the curvilinear structures of interest. These ranges are easily established by image inspection. To estimate local orientation (assumed known in Equation 4) we expand our filter bank with kernel replicas rotated by $\theta \in [0, 2\pi)$, $\{\theta^{(i)},\ i = 1, \ldots, N_\theta\}$. Finally, we account for fragmented (point-like) structures by adopting a suitable range for the "memory" parameter σ_1, $\{\sigma_1^{(i)},\ i = 1, \ldots, N_{\sigma_1}\}$. Figure 3 shows some of the resulting convolutional tubularity kernels used in our experiments.

Once the filter bank has been precomputed, scale, curvature, rotation, and elongation invariance is obtained by maximising the tubularity measure $I_{\mathbf{ww}}(x, y; \sigma_1, \sigma_2, k, \theta)$ for each pixel $(x, y) \in \mathcal{D}(I)$ across the pre-defined parameter space $\{(\sigma_1, \sigma_2, k, \theta)^{(j)}, j = 1, \ldots, N_{\sigma_1} + N_{\sigma_2} + N_k + N_\theta\}$:

$$\left(\sigma_1, \sigma_2, k, \theta\right)^{*} = \underset{\sigma_1, \sigma_2, k, \theta}{\arg\max}\, I_{\mathbf{ww}}\left(x, y; \sigma_1, \sigma_2, k, \theta\right). \tag{A.7}$$

Notice that, by definition, the maximum value of $I_{\mathbf{ww}}(x, y; \sigma_1, \sigma_2, k, \theta)$ at each pixel across the parameter space corresponds to the measure of tubularity we defined in Equation A.4. Since our tubularity measure is based on contrast, its value, computed on tubular objects, should be always either positive or negative, and its sign depends on whether the inside or the outside region is brighter. When $I_{\mathbf{ww}}$ should be positive, for instance, negative local tubularity responses are likely due to a non-target objects and they should be discarded. To this aim, we apply soft thresholding such that $I_{\mathbf{ww}}^{st} = \max\left(I_{\mathbf{ww}}, 0\right)$.

Several confocal microscopy images in our data set show significant intra- and inter-image contrast variability (Figure 1), due to illumination and the reflectance of the ocular tissue. Hence, local tubularity estimation tends to be biased towards high contrast, penalising thin and poorly contrasted corneal nerve fibres. To rectify this problem and achieve contrast invariance, we introduce a contrast normalization term and a parameter $\alpha \in \mathbb{R}$ in the proposed ridge detector, emphasising the response of low-contrast structures:

$$SCIRD\Big(x, y; \{\sigma_1, \sigma_2, k, \theta\}^*\Big) = \frac{I^{st}_{\mathbf{ww}}(x, y; \{\sigma_1, \sigma_2, k, \theta\}^*)}{1 + \alpha\, I^C(x, y)}, \quad \text{(A.8)}$$

where

$$I^C(x, y) = \frac{1}{(N+1)^2} \sum_{i=x-\frac{N}{2}}^{x+\frac{N}{2}} \sum_{j=y-\frac{N}{2}}^{y+\frac{N}{2}} \max_{\sigma_2} \left\| \nabla I(i, j; \sigma_2) \right\|_2 \quad \text{(A.9)}$$

is the adopted contrast measure based on multiscale gradient magnitude estimation averaged on a patch $(N + 1) \times (N + 1)$ around pixel $(x, y) \in \mathcal{D}(I)$. Notice, N is not a free parameter, but the width (and height) of the largest filter in the SCIRD filter bank ($N = 8\sigma_{2max}$).

References

Annunziata, R., Garzelli, A., Ballerini, L., Mecocci, A., Trucco, E., 2015a. Leveraging multiscale hessian-based enhancement with a novel exudate inpainting technique for retinal vessel segmentation. IEEE Journal of Biomedical and Health Informatics (in press) .

Annunziata, R., Kheirkhah, A., Aggarwal, S., Cavalcanti, B.M., Hamrah, P., Trucco, E., 2014. Tortuosity classification of corneal nerves images using a multiple-scale-multiple-window approach, in: Proceedings of the Ophthalmic Medical Image Analysis (OMIA) First International Workshop, MICCAI 2014, pp. 113–120.

Annunziata, R., Kheirkhah, A., Hamrah, P., Trucco, E., 2015b. Boosting handcrafted features for curvilinear structure segmentation by learning context filters, in: Medical Image Computing and Computer Assisted Interventions (MICCAI), LNCS. pp. 596–603.

Annunziata, R., Kheirkhah, A., Hamrah, P., Trucco, E., 2015c. Scale and curvature invariant ridge detector for tortuous and fragmented structures, in: Medical Image Computing and Computer Assisted Interventions (MICCAI), LNCS. pp. 588–595.

Becker, C., Rigamonti, R., Lepetit, V., Fua, P., 2013. Supervised feature learning for curvilinear structure segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, pp. 526–533.

Bishop, C.M., et al., 2006. Pattern recognition and machine learning. volume 1. springer New York.

Bogunović, H., Pozo, J.M., Cárdenes, R., Villa-Uriol, M.C., Blanc, R., Piotin, M., Frangi, A.F., 2012. Automated landmarking and geometric characterization of the carotid siphon. Medical image analysis 16, 889–903.

Bribiesca, E., 2013. A measure of tortuosity based on chain coding. Pattern Recognition 46, 716 – 724.

Budai, A., Hornegger, J., Michelson, G., 2009. Multiscale approach for blood vessel segmentation on retinal fundus images. Investigative Ophthalmology & Visual Science 50, 325–325.

Bullitt, E., Ewend, M.G., Aylward, S., Lin, W., Gerig, G., Joshi, S., Jung, I., Muller, K., Smith, J.K., 2004. Abnormal vessel tortuosity as a marker of treatment response of malignant gliomas: preliminary report. Technology in cancer research & treatment 3, 577–584.

Bullitt, E., Gerig, G., Pizer, S.M., Lin, W., Aylward, S.R., 2003. Measuring tortuosity of the intracerebral vasculature from mra images. Medical Imaging, IEEE Transactions on 22, 1163–1171.

Bullitt, E., Muller, K.E., Jung, I., Lin, W., Aylward, S., 2005. Analyzing attributes of vessel populations. Medical image analysis 9, 39–49.

Cairney, J., 1924. Tortuosity of the cervical segment of the internal carotid artery. Journal of Anatomy 1, 87–96.

Cheung, C.Y.l., Zheng, Y., Hsu, W., Lee, M.L., Lau, Q.P., Mitchell, P., Wang, J.J., Klein, R., Wong, T.Y., 2011. Retinal vascular tortuosity, blood pressure, and cardiovascular risk factors. Ophthalmology 118, 812–818.

Coates, A., Ng, A.Y., 2012. Learning feature representations with k-means, in: Neural Networks: Tricks of the Trade. Springer, pp. 561–580.

Coulston, J., Baigent, A., Selvachandran, H., Jones, S., Torella, F., Fisher, R., 2014. The impact of endovascular aneurysm repair on aortoiliac tortuosity and its use as a predictor of iliac limb complications. Journal of Vascular Surgery 60, 585 – 589.

Criminisi, A., Shotton, J., Konukoglu, E., 2012. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Foundations and Trends® in Computer Graphics and Vision 7, 81–227.

Cruzat, A., Witkin, D., Baniasadi, N., Zheng, L., Ciolino, J.B., Jurkunas, U.V., Chodosh, J., Pavan-Langston, D., Dana, R., Hamrah, P., 2011. Inflammation and the nervous system: The connection in the cornea in patients with infectious keratitis. Investigative Ophthalmology Visual Science 52, 5136–5143.

Dougherty, G., Varro, J., 2000. A quantitative index for the measurement of the tortuosity of blood vessels. Medical engineering & physics 22, 567–574.

Dunn, G., 1989. Design and analysis of reliability studies: The statistical evaluation of measurement errors. Edward Arnold Publishers.

Edington, G., 1901. Tortuosity of both internal carotic arteries. British Medical Journal 2, 1526–7.

Edwards, K., Pritchard, N., Vagenas, D., Russell, A., Malik, R.A., Efron, N., 2014. Standardizing corneal nerve fibre length for nerve tortuosity increases its association with measures of diabetic neuropathy. Diabetic Medicine 31, 1205–1209.

Eleid, M.F., Guddeti, R.R., Tweet, M.S., Lerman, A., Singh, M., Best, P.J., Vrtiska, T.J., Prasad, M., Rihal, C.S., Hayes, S.N., et al., 2014. Coronary artery tortuosity in spontaneous coronary artery dissection angiographic characteristics and clinical implications. Circulation: Cardiovascular Interventions 7, 656–662.

Farabet, C., Couprie, C., Najman, L., LeCun, Y., 2013. Learning hierarchical features for scene labeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35, 1915–1929.

Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A., 1998. Multiscale vessel enhancement filtering, in: Medical Image Computing and Computer-Assisted Interventation (MICCAI). Springer, pp. 130–137.

Franken, R., el Morabit, A., de Waard, V., Timmermans, J., Scholte, A.J., van den Berg, M.P., Marquering, H., Planken, N.R., Zwinderman, A.H., Mulder, B.J., et al., 2015. Increased aortic tortuosity indicates a more severe aortic phenotype in adults with marfan syndrome. International journal of cardiology 194, 7–12.

Grisan, E., Foracchia, M., Ruggeri, A., 2008. A novel method for the automatic grading of retinal vessel tortuosity. IEEE Trans. Med. Imaging 27, 310–319.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. The Journal of Machine Learning Research 3, 1157–1182.

Hamrah, P., Cruzat, A., Dastjerdi, M.H., Prss, H., Zheng, L., Shahatit, B.M., Bayhan, H.A., Dana, R., Pavan-Langston, D., 2013. Unilateral herpes zoster ophthalmicus results in bilateral corneal nerve alteration: An in vivo confocal microscopy study. Ophthalmology 120, 40 – 47.

Hamrah, P., Cruzat, A., Dastjerdi, M.H., Zheng, L., Shahatit, B.M., Bayhan, H.A., Dana, R., Pavan-Langston, D., 2010. Corneal sensation and subbasal nerve alterations in patients with herpes simplex keratitis: An in vivo confocal microscopy study. Ophthalmology 117, 1930 – 1936.

Hannink, J., Duits, R., Bekkers, E., 2014. Crossing-preserving multiscale vesselness, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer, pp. 603–610.

Hart, W.E., Goldbaum, M., Côté, B., Kube, P., Nelson, M.R., 1999. Measurement and classification of retinal vascular tortuosity. International journal of medical informatics 53, 239–252.

Heneghan, C., Flynn, J., OKeefe, M., Cahill, M., 2002. Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis. Medical image analysis 6, 407–429.

Honnorat, N., Vaillant, R., Paragios, N., 2011. Graph-based geometric-iconic guide-wire tracking, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011. Springer, pp. 9–16.

Ikram, M., Ong, Y., Cheung, C., Wong, T., 2013. Retinal vascular caliber measurements: Clinical significance, current knowledge and future perspectives.

Ophthalmologica .

Ji, J., Shimony, J., Gao, F., McKinstry, R.C., Gutmann, D.H., 2013. Optic nerve tortuosity in individuals with neurofibromatosis type 1. Pediatric radiology 43, 1336.

Joshi, V., Reinhardt, J.M., Abràmoff, M.D., 2010. Automated measurement of retinal blood vessel tortuosity, in: SPIE Medical Imaging, International Society for Optics and Photonics. pp. 76243A–76243A.

Kallinikos, P., Berhanu, M., O'Donnell, C., Boulton, A.J., Efron, N., Malik, R.A., 2004. Corneal nerve tortuosity in diabetic patients with neuropathy. Investigative ophthalmology & visual science 45, 418–422.

Koprowski, R., Teper, S.J., Weglarz, B., Wylegala, E., Krejca, M., Wróbel, Z., 2012. Fully automatic algorithm for the analysis of vessels in the angiographic image of the eye fundus. Biomedical Engineering Online 11.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Kurbanyan, K., Hoesl, L., Schrems, W., Hamrah, P., 2012. Corneal nerve alterations in acute acanthamoeba and fungal keratitis: an in vivo confocal microscopy study. Eye 26, 126–132.

Lagali, N., Poletti, E., Patel, D.V., McGhee, C.N.J., Hamrah, P., Kheirkhah, A., Tavakoli, M., Petropoulos, I.N., Malik, R.A., Utheim, T.P., Zhivov, A., Stachs, O., Falke, K., Peschel, S., Guthoff, R., Chao, C., Golebiowski, B., Stapleton, F., Ruggeri, A., 2015. Focused tortuosity definitions based on expert clinical assessment of corneal subbasal nervesexpert assessment of corneal nerve tortuosity. Investigative Ophthalmology Visual Science 56, 5102–5109.

Law, M.W., Chung, A.C., 2008. Three dimensional curvilinear structure detection using optimally oriented flux, in: Computer Vision–ECCV 2008. Springer, pp. 368–382.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., Nowe, A., 2012. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 9, 1106–1119.

Lesage, D., Angelini, E.D., Bloch, I., Funka-Lea, G., 2009. A review of 3d vessel lumen segmentation techniques: Models, features and extraction schemes. Medical image analysis 13, 819–845.

Lin, J.K., Dayan, P., 1999. Curved gaussian models with application to the modeling of foreign exchange rates, in: Computational Finance, MIT Press.

Lisowska, A., Annunziata, R., Loh, G., Karl, D., Trucco, E., 2014. An experimental assessment of five indices of retinal vessel tortuosity with the ret-tort public dataset, in: Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, pp. 5414–5417.

Longmuir, S.Q., Mathews, K.D., Longmuir, R.A., Joshi, V., Olson, R.J., Abràmoff, M.D., 2010. Retinal arterial but not venous tortuosity correlates with facioscapulohumeral muscular dystrophy severity. Journal of American Association for Pediatric Ophthalmology and Strabismus 14, 240–243.

Martinez-Perez, M.E., Hughes, A.D., Thom, S.A., Bharath, A.A., Parker, K.H., 2007. Segmentation of blood vessels from red-free and fluorescein retinal images. Medical image analysis 11, 47–61.

Maude, R.J., Ahmed, B.W., Rahman, A.H., Rahman, R., Majumder, M.I., Menezes, D.B., Sayeed, A.A., Hughes, L., MacGillivray, T.J., Borooah, S., et al., 2014. Retinal changes in visceral leishmaniasis by retinal photography. BMC infectious diseases 14, 527.

Muraoka, Y., Tsujikawa, A., Kumagai, K., Akagi-Kurashige, Y., Ogino, K., Murakami, T., Miyamoto, K., Yoshimura, N., 2014. Retinal vessel tortuosity associated with central retinal vein occlusion: An optical coherence tomography studytortuous retinal vessels in crvo. Investigative ophthalmology & visual science 55, 134–141.

Oliveira-Soto, L., Efron, N., 2001. Morphology of corneal nerves using confocal microscopy. Cornea 20, 374–384.

Owen, C.G., Newsom, R.S., Rudnicka, A.R., Barman, S.A., Woodward, E.G., Ellis, T.J., 2008. Diabetes and the tortuosity of vessels of the bulbar conjunctiva. Ophthalmology 115, e27–e32.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27, 1226–1238.

Rigamonti, R., Lepetit, V., 2012. Accurate and efficient linear structure segmentation by leveraging ad hoc features with learned filters, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012. Springer, pp. 189–197.

Santamaría-Pang, A., Colbert, C., Saggau, P., Kakadiaris, I.A., 2007. Automatic centerline extraction of irregular tubular structures using probability volumes from multiphoton imaging, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007. Springer, pp. 486–494.

Sasongko, M., Wong, T., Nguyen, T., Cheung, C., Shaw, J., Wang, J., 2011. Retinal vascular tortuosity in persons with diabetes and diabetic retinopathy. Diabetologia 54, 2409–2416.

Sasongko, M.B., Wong, T.Y., Nguyen, T.T., Cheung, C.Y., Shaw, J.E., Kawasaki, R., Lamoureux, E.L., Wang, J.J., 2015. Retinal vessel tortuosity and its relation to traditional and novel vascular risk markers in persons with diabetes. Current eye research , 1–7.

Scarpa, F., Zheng, X., Ohashi, Y., Ruggeri, A., 2011. Automatic evaluation of corneal nerve tortuosity in images from in vivo confocal microscopy. Investigative ophthalmology & visual science 52, 6404–6408.

Schneider, M., Hirsch, S., Weber, B., Székely, G., Menze, B.H., 2015. Joint 3-d vessel segmentation and centerline extraction using oblique hough forests with steerable filters. Medical image analysis 19, 220–249.

Sironi, A., Lepetit, V., Fua, P., 2014. Multiscale centerline detection by learning a scale-space distance transform, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE. pp. 2697–2704.

Soares, J.V., Leandro, J.J., Cesar Jr, R.M., Jelinek, H.F., Cree, M.J., 2006. Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. Medical Imaging, IEEE Transactions on 25, 1214–1222.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45, 427–437.

Trucco, E., Azegrouz, H., Dhillon, B., 2010. Modeling the tortuosity of retinal vessels: does caliber play a role? Biomedical Engineering, IEEE Transactions on 57, 2239–2247.

Tu, Z., Bai, X., 2010. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. IEEE TPAMI .

Wilson, C.M., Cocker, K.D., Moseley, M.J., Paterson, C., Clay, S.T., Schulenburg, W.E., Mills, M.D., Ells, A.L., Parker, K.H., Quinn, G.E., et al., 2008. Computerized analysis of retinal vessel width and tortuosity in premature infants. Investigative ophthalmology & visual science 49, 3577–3585.

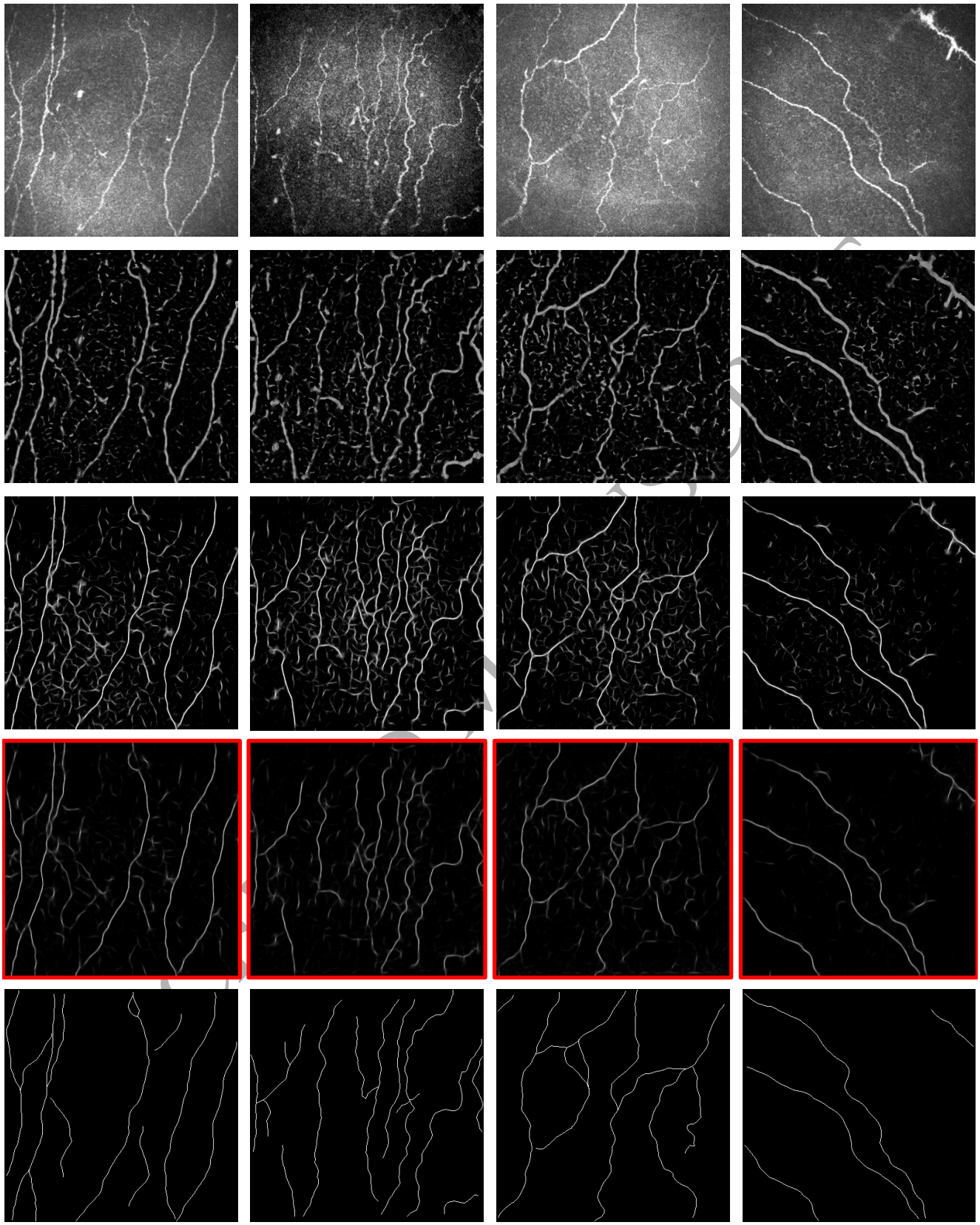


Figure 10: Examples from segmentation experiments. First row: original images showing corneal nerve fibres in the subbasal layer. From second to fourth row: probability (i.e., tubularity) maps obtained by combining OOF with learned *appearance* filters (Rigamonti and Lepetit, 2012), combining OOF with *single*-range context filters (Annunziata et al., 2015b) and combining SCIRD with *multi*-range learned context filters (proposed approach - results highlighted in red). Last row: manual tracing used as reference ground truth. The proposed segmentation approach shows a significant reduction of false positives detected in background regions far from target structures by other methods and a better connectivity of structures appearing fragmented.